The international journal of science / 9 January 2020

nature

UNEQUAL OPPORTUNITIES

Disparities in childhood education and nutrition in low- and middle-income countries

Driving force
The Piezo proteins that help cells to sense touch and pressure

Paper trail
A checklist to assess the integrity of published articles

Source data
Expanded origins for repeating fast radio bursts

Vol. 577, No. 7789
nature.com

The international journal of science / 9 January 2020

nature

Injecting new life into an old vaccine

The BCG vaccine is nearing its centenary. A new delivery method could protect more people from one of the world's biggest killers.

Tuberculosis is a leading cause of death worldwide and the biggest killer among people with HIV. Despite widespread efforts to boost vaccination and treatment, each year around 10 million people become ill from the disease, and in 2018 some 1.5 million people died, according to the World Health Organization (WHO). Astonishingly, the only licensed tuberculosis vaccine – BCG (Bacillus Calmette–Guérin), made from a weakened strain of the disease-causing bacterium, *Mycobacterium tuberculosis* – has been in use for almost a century. Although BCG ably protects infants from systemic infection, it does less well against the lung disease that tuberculosis causes in adolescents and adults.

Newer anti-tuberculosis medicines have been developed, and more drugs and vaccines are in clinical trials. Nonetheless, the two most powerful drugs currently in use are not effective against multidrug-resistant strains of *M. tuberculosis*.

A new study in non-human primates (P. A. Darrah *et al. Nature* **577**, 95–102; 2020) could breathe new life into the BCG vaccine. The key advance it reports is that delivery of the vaccine intravenously – directly into the bloodstream – rather than into skin or muscle, boosted protection against tuberculosis, particularly in the lungs. The translation of those findings into clinical practice is still some way off, but the pay-off could be huge.

In the study, researchers in the United States compared the results when BCG was delivered in three different ways to separate groups of macaque monkeys (*Macaca mulatta*): intravenously; by injection into the skin; or through an aerosol spray. The team found that intravenous administration yielded more T cells – a type of immune cell – that responded to *M. tuberculosis* than did either of the other routes. Moreover, these T cells were found throughout the animals' lungs.

Six months later, some of the macaques were exposed to *M. tuberculosis*. Of those who received the vaccine intravenously, nine out of ten animals were highly protected against disease and six showed no detectable signs of infection. By contrast, most of the macaques vaccinated by other routes exhibited hallmarks of infection.

The results were a surprise: it was thought that intravenous delivery of the vaccine might limit the severity of disease, but not prevent infection altogether. Instead, intravenous BCG delivery seems to trigger the establishment of a population of T cells in the lungs that can rapidly neutralize *M. tuberculosis*.

> "The risks of injecting a live bacterium into the bloodstream need to be better understood."

If these results hold up in subsequent studies, this would represent significant promise for tuberculosis control. A cheap and already widely used vaccine could confer greater resistance to disease with only a relatively small change to the way in which it is administered. But there are a number of hurdles to overcome.

The first is safety. As with most vaccines, BCG contains live bacteria that have been weakened, making them unable to infect people with robust immune systems. But the risks of injecting a live bacterium directly into the bloodstream – as opposed to injecting it into the skin – need to be better understood. If, for some reason, a vaccine became contaminated by other bacteria or contained *M. tuberculosis* capable of causing disease, an infection could quickly become life-threatening. Researchers must complete careful and exhaustive safety tests before carrying the technique over into clinical trials in humans.

Further obstacles

If the new approach is found to be safe, testing intravenous administration of the vaccine and then delivering it in people is likely to present further obstacles. It is more difficult to administer a vaccine intravenously than to give a quick jab into skin or muscle – and doing so will require extra funds to train medical personnel in countries where underfunded public-health systems already struggle to provide adequate primary health care.

This lack of funding, and the fact that tuberculosis mostly affects people in poorer countries, is one reason why drug companies have not developed new treatments. However, newer players – particularly not-for-profit research groups – are changing drug discovery. One leading vaccine candidate, M72/AS01$_E$, developed through a collaboration between a non-profit organization and a pharmaceutical company, protected adults with latent *M. tuberculosis* infection from going on to develop active tuberculosis for at least three years (D. R. Tait *et. al. N. Engl. J. Med.* **381**, 2429–2439; 2019). And last summer, the US Food and Drug Administration approved a drug – also developed through not-for-profit research – as part of a combination to treat an extreme form of multidrug-resistant tuberculosis (*Nature* http://doi.org/dhvs; 2019). In all, the WHO says that there are 23 drugs, several combination therapies and 14 vaccine candidates in clinical trials.

Intravenous vaccines are being developed in other areas, too. One malaria vaccine called PfSPZ is due to be tested in a large clinical trial in 2,100 people this year in Bioko, Equatorial Guinea (*Nature* http://doi.org/gfzjq2; 2019).

Existing tuberculosis treatments saved between 53 million and 64 million lives between 2000 and 2018. But it is unacceptable that so many people still fall ill with the disease. Under the United Nations Sustainable Development Goals, tuberculosis deaths need to be reduced by 90% by 2030, and BCG vaccination will be a crucial part of this effort. Universal vaccination in infants has been recommended in more than 150 countries – including many in Africa and southeast Asia, which have been particularly hard hit by the disease. A potential means of boosting a vaccine's power warrants the world's attention.

© 2020 Springer Nature Limited. All rights reserved.

World view



By Heidi Rehm

New Year's resolution: one meeting-free week a quarter

Protect time for deep thinking — it is crucial for productivity in a world of constant interaction, urges Heidi Rehm.

Last December, I analysed how much time I'd devoted to meetings and conference calls during the previous three months. Discounting the days when I had been travelling or at conferences and offsite meetings, I had averaged 8.5 hours a day. This year, I have a plan to protect work time: I'm scheduling one week a quarter during which meetings and conference calls are not allowed in my research group. So far, eight other groups have joined in this plan, and we expect others to come on board. More than 500 people expressed support for my call on Twitter.

I split my time between academic and clinical work, discovering the causes of rare disease, implementing genomics in health care, and building tools to support genomic medicine. I usually start work at 6 a.m., which allows me 2 hours of thoughtful work time: that includes following up on e-mails that demand special attention, and preparing for lectures and meetings, which typically begin at 8 a.m. I attend to Slack and e-mail communications in the evening. Writing and editing, preparing new presentations, and other more-substantive projects are for weekends and flights. That leaves little time for major projects, writing grants or doing deep critical or innovative thinking. This is what I want meeting-free weeks for.

I stumbled across the idea by accident. Several years ago, my holiday plans fell through, but I left the block on my calendar, preventing any meetings from being scheduled. When the week arrived, I confessed that I planned to spend the time at home with my children, who were on holiday from school. But with my younger son at basketball camp and my older daughter determined to be independent, I found myself free to dive into work with little disruption. I caught up on the long list of tasks that had piled up, particularly writing and editing. I got my e-mail inbox down to zero by the end of the week, and I also had time to watch films with my family in the evenings. It was magnificent!

I began doing this at least one week a year. Members of my lab quickly worked out what was going on, so I stopped pretending and made it clearer which weeks were actual holidays (when I really am mostly off the grid), and when it was 'Heidi's working at home without meetings' week. They respected my plan because they knew that I would get the things done that they needed from me.

Other groups at the Broad Institute in Cambridge, Massachusetts, including the Genomics Platform, the Data Sciences Platform and the lab of Daniel MacArthur, began implementing one meeting-free week per quarter. It has been extremely positive, with people reporting being highly productive and able to dive into larger, sometimes daunting, projects that require longer stretches of dedicated attention. However, because more than three-quarters of my meetings and calls are outside the institute, those weeks still had at least 30 hours scheduled.

> "Block out the weeks on the calendar immediately."

So, late last year, I extended the call to the broader genomics community. I crafted a proposal that avoids clashes with major genomics meetings and alternates which week of the month is free, to avoid repeated conflicts with monthly meetings (go.nature.com/2en531y). Several groups have since told me that they are working to implement the practice in their own programmes and labs.

My suggestion is to block out the weeks on the calendar immediately. Next, start cancelling and rescheduling meetings booked for those weeks. Standing weekly meetings can almost certainly be shed, as can many monthly ones, which will be disrupted only once a year. It is OK to keep very difficult-to-schedule meetings. Once the week arrives, ad hoc meetings are allowed, at the discretion of the group head, but these should have the specific purpose of accomplishing the work planned for the meeting-free week.

Although I have not found studies comparing approaches to curbing meetings, several surveys suggest that we are overly burdened with them. A survey of 182 senior managers across a range of industries found that about two-thirds felt that meetings cut into deep thinking and kept them from completing their own work (L. A. Perlow *et al. Harvard Business Rev.* 62–69; July–August 2017).

Of course, four weeks per year is a paltry amount of time for people whose job it is, essentially, to be innovative. Researchers, research groups and work places need to brainstorm their own systems to prevent meetings from degrading productivity, rigour and innovation. Some people carve out one day per week without meetings, others a period of time each day. Some companies arrange company-wide protected time. I encourage every person and programme to find a system that works for them and stick to it.

In addition to taking time out from meetings, notice other barriers to productivity, perhaps e-mail or social media. Create periods of time that are free from these distractions or give them up entirely. Take a sabbatical if a change of environment is needed.

Personal strategies aside, with today's extensive, globe-spanning collaborations, it makes sense to develop an international, coordinated effort across programmes and institutions, and find four weeks a year to align meeting-free time globally.

I encourage everyone to enjoy these quarterly fasts. Let's report back to one another at the end of the year. Perhaps it will help to bring the research community together more deeply.

Heidi Rehm is the chief genomics officer at Massachusetts General Hospital and the medical director at the Broad Institute's Clinical Research Sequencing Platform. e-mail: HREHM@mgh.harvard.edu

KATE FLANAGAN

© 2020 Springer Nature Limited. All rights reserved.

News in brief

IRAN TAKES PIVOTAL NUCLEAR STEP

Iran has announced that it will no longer abide by any restrictions on its uranium-enrichment programme, prompting fears that the 2015 deal to limit the country's nuclear activities is effectively dead. Experts say that if Iran resumes enrichment at full throttle, it could be capable of creating a nuclear bomb in one year.

The step is widely seen as a response to the 3 January US drone strike on Baghdad that killed visiting Iranian military commander Qasem Soleimani, the architect of Iran's military operations around the world.

Iran struck the 2015 nuclear deal, the Joint Comprehensive Plan of Action (JCPOA), with six global powers on the condition that punishing international sanctions would be lifted. It agreed to drastic cuts in its ability to produce enriched uranium and plutonium, which can be used to build a nuclear bomb. In particular, the deal limited the number of centrifuges Iran could use to separate uranium isotopes, and the amount and purity of uranium-235 the country could stockpile.

The United States unilaterally pulled out of the JCPOA in 2018 and resumed harsh economic sanctions against the country. In May 2019, Iran announced partial removal of its restrictions. On 5 January, foreign minister Mohammad Javad Zarif tweeted that there will be no more limits on the numbers of centrifuges. He added that Iran will continue to allow international nuclear inspections and that it will revert to full compliance with the deal if the other parties do the same.



SCI-HUB PROBED FOR ALLEGED LINKS TO RUSSIAN INTELLIGENCE

The operator of Sci-Hub, a website that provides people worldwide with illicit access to a large volume of copyright-protected academic literature, is reportedly being investigated by the US Department of Justice (DOJ) on suspicion of stealing US military secrets.

The agency suspects Kazakhstan-born Alexandra Elbakyan, who created Sci-Hub in 2011, of running the pirate site with silent approval from the Russian government and, possibly, with support from military intelligence. *The Washington Post* reported the allegations on 19 December, citing anonymous sources close to the DOJ. The DOJ told *Nature* that it could neither confirm nor deny the existence of an ongoing investigation.

Sci-Hub allegedly operates from servers in Russia. Critics say that Elbakyan, a computer scientist whose whereabouts are unknown, uses advanced hacking and phishing practices to acquire login details illegally from libraries and personal subscribers. Elbakyan previously told *Nature* that she runs Sci-Hub alone as a private enterprise, with financial help from anonymous donors. She did not respond to *Nature*'s request for comment on her alleged links to Russian intelligence.

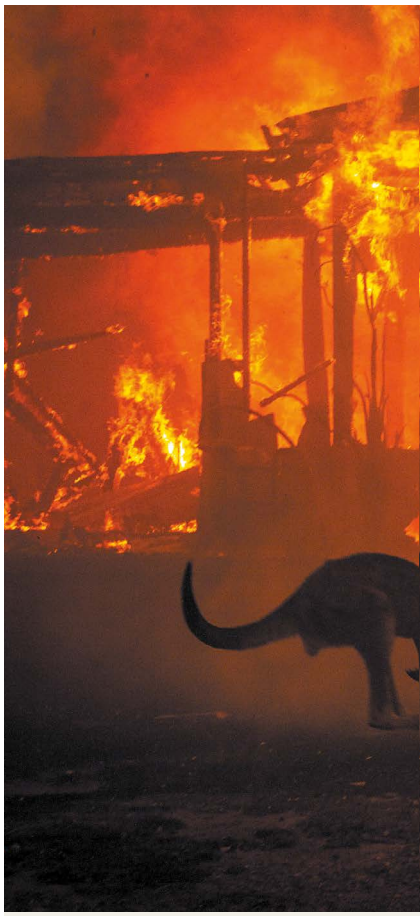
UK GOVERNMENT CALL FOR SCIENCE 'WEIRDOS' PROMPTS CAUTION

Researchers have reacted with surprise to a bizarrely worded job advertisement posted by a senior UK government adviser. The notice calls for scientists, mathematicians and "super-talented weirdos" to work for the prime minister, and cites several scientific papers.

Dominic Cummings (pictured), a political strategist who is chief special adviser to Prime Minister Boris Johnson, posted the advertisement on his personal website on 2 January. Cummings, known for his strong and sometimes controversial views on science, says that he is seeking "data scientists, project managers, policy experts, assorted weirdos", as well as "unusual" mathematicians, physicists and economists, to work in the prime minister's office in Downing Street. The post attracted attention for its brash language and odd requirements, such as that applicants show excitement about a slew of specified scientific ideas.

Researchers, including the authors of some of the cited papers, welcomed a focus on data-driven techniques and scientific skills from the top level of government, but cautioned against oversimplifying how science is applied to policymaking.





Bush fires 'twice the size of Belgium' rage across Australia

L TO R: ALEXANDER KRASSOTKIN/WIKIMEDIA COMMONS; DANIEL LEAL-OLIVAS/AFP/GETTY; MATTHEW ABBOTT/NYT/REDUX/EYEVINE; BRENDON THORNE/GETTY

© 2020 Springer Nature Limited. All rights reserved.

Nature Briefing

Sign up to get essential science news, opinion and analysis delivered to your inbox daily. Visit go.nature.com/newsletter



Nearly half a billion native animals are estimated to have perished in bush fires in New South Wales, the Australian state most affected by the ongoing disastrous fires, and where this image was captured.

The total number of animals killed is likely to be much higher across the country, where conflagrations have been burning in multiple states since September.

The fires have so far burnt through at least six million hectares of land — more than twice the size of Belgium — across five states. In three states alone — New South Wales, Victoria and South Australia — fires have claimed at least 24 human lives and destroyed more than 1,800 homes.

The scale of the disaster prompted the federal government to call on the country's defence forces to help fire-affected communities.

On 3 January, residents of coastal towns in New South Wales and Victoria were told to leave ahead of worsening conditions, leading to one of the country's largest mass evacuations on record. About 160 fires were burning across those states as *Nature* went to press.

RUMOURED CHANGES TO US GOVERNMENT OPEN-ACCESS POLICY

A rumour that the White House is considering a policy that would make all federally funded studies free to read on publication has fuelled a debate between scientists who favour open access and publishers of subscription journals.

According to the widely discussed rumour, whose origin is unclear, the administration of President Donald Trump is drafting an executive order that would force the change in publishing practices.

On 18 December, two groups that represent publishers — the Association of American Publishers in Washington DC and the International Association of Scientific, Technical, and Medical Publishers in Oxford, UK — sent letters to the US government opposing any such policy. It would hinder the peer-review process, stifle innovation and tip the publishing business into chaos, they wrote.

But several scientists who advocate open access countered the publishers' arguments. "I welcome the rumored policy," tweeted John Wilbanks, chief commons officer at the non-profit research organization Sage Bionetworks in Seattle, Washington. "I work at a *really well funded* non-traditional research organization. We still can't afford journal access subscriptions."

Kristina Baum, a spokesperson for the White House Office of Science and Technology Policy, declined to comment on the rumour.

The rumour follows an effort led by European funders, called Plan S, that will require that research they fund be made open access on publication, with liberal licensing terms.



HEAD OF ANCIENT-DNA LAB SACKED FOR MISCONDUCT

The University of Adelaide has fired Alan Cooper, the high-profile leader of its Australian Centre for Ancient DNA, for "serious misconduct". His dismissal follows allegations that Cooper bullied staff and students, and an investigation into the 'culture' of the centre.

Cooper is a major figure in the field of ancient DNA, and has charted the migrations of prehistoric people and their domestic animals around the globe.

In a statement, the university said it would not comment on the circumstances that led to Cooper's dismissal, but said that the move did not relate to the quality or integrity of his research or any other research conducted at the centre.

"The university takes seriously its commitment to the welfare of students and staff," it also said.

In response to his dismissal, Cooper told *Nature* that he rejects the allegation that he was a bully. "I work at the highest international levels, and want my students and staff to do the same. I've occasionally been too blunt in my language and actions, and regret this — but it was never bullying," he says. Cooper says that he was never warned about his conduct before action was taken against him, and that he was never asked to attend management courses.

© 2020 Springer Nature Limited. All rights reserved.

News in focus



STEPHANE DE SAKUTIN/AFP/GETTY

The devastating fire at Notre-Dame de Paris last April destroyed the cathedral's roof and spire.

THE HUGE SCIENTIFIC EFFORT TO STUDY NOTRE-DAME'S ASHES

Last year's fire at Paris's beloved cathedral shocked the world. Now, researchers are making use of the unprecedented opportunity to study its materials.

By Philip Ball

The fire that destroyed large sections of the iconic cathedral Notre-Dame de Paris last April was a national tragedy. Now, months on, scientists with the French national research organization CNRS are embarking on a multi-million-euro effort to study the 850-year-old building and its materials with the goal of illuminating how it was constructed. With unprecedented access to the cathedral's fabric — including timber, metalwork and the building's foundations — in the wake of the fire, scientists also hope that their work will arm them with information to help the restoration.

The research could "write a new page in the history of Notre-Dame, because there are currently many grey areas," says Yves Gallet, a historian of Gothic architecture at the University of Bordeaux-Montaigne, who is in charge of a 30-strong research team investigating the masonry.

Construction of the cathedral, considered one of the finest examples of the French Gothic style, began in the twelfth century. The structure was modified in the Middle Ages and extensively restored in the nineteenth century. But it has been the subject of surprisingly little scientific research compared with other Gothic monuments in France and elsewhere, says Martine Regert, a biomolecular archaeologist at the CNRS's CEPAM centre for the study of historical cultures and environments in Nice, who is one of the project's leaders. Many questions remain about the structure, such as which sections are medieval and whether architect Eugène Viollet-Le-Duc, who supervised the nineteenth-century renovation, reused some of the older materials, says Regert.

The fire on 15 April, possibly caused by an electrical fault, destroyed the cathedral's roof and spire, and caused part of its vaulted

© 2020 Springer Nature Limited. All rights reserved.

ceiling to collapse. The walls still stand, and the building will eventually be restored – although this is likely to take longer than the ambitious five years initially forecast, and is set to cost hundreds of millions of euros.

But, until then, the interior of the building holds piles of debris: fallen stonework, burnt timbers and damaged metal artefacts, all now available for scientific study. The absence of tourists might also make it possible to use radar to probe the foundations, which have been little investigated before. Even some parts of the structure that were largely undamaged are now more accessible for inspection, says metals specialist Philippe Dillmann of the CNRS Laboratory for Archaeomaterials and Alteration Forecasting in Gif-sur-Yvette, who is coordinating the project with Regert.

Architectural investigations

The CNRS project will focus on seven topics: masonry, wood, metalwork, glass, acoustics, digital data collection and anthropology. In all, the effort will involve more than 100 researchers in 25 laboratories and will last for 6 years.

Gallet's team will study Notre-Dame's stones to identify the quarries that supplied them and "reconstruct the supply networks and the economy of the site". Studying the mortar used to bind the stones together could reveal how different compositions were used for the various structural elements – vaulting, walls and flying buttresses. A better knowledge of the historical materials could inform choices made in restoration, says Gallet.

The team will also analyse weaknesses in the remaining structure caused by the high temperatures of the fire, and by the fall of masonry and the water used to extinguish the flames. A radar study will determine how solid the foundations are before restorers erect scaffolding to allow them to dismantle the unstable remnants of the nineteenth-century spire.

Meanwhile, a team of about 50 will focus on Notre-Dame's famous woodwork – especially the timbers in the roof space above the vaults, dubbed the forest – which has either burnt away or lies charred in the nave. These blackened remains could be tremendously valuable to researchers.

"The burnt structure constitutes a gigantic laboratory for archaeology," says dendrologist Alexa Dufraisse of the National Museum of Natural History in Paris, who will lead the multidisciplinary wood team.

"Wood is an extraordinary source of information," says Regert. Initial observations have confirmed that the 'forest' is made of oak, but studies will pinpoint the exact species used and give researchers clues about the techniques and tools of medieval timber construction.

Tree-ring dating of timber beams could reveal the year and location in which the trees were felled, filling in gaps in knowledge about the sequence of construction. "Each tree records within its tissues the environment in which it has grown," says Dufraisse. This kind of study "could never have been conducted without the destruction of the structure by fire", she says.

In particular, says Regert, the wood is a climate archive. "Isotopic analyses of oxygen and carbon in the rings make it possible to determine the temperature and rainfall over time," she says. The trees used in Notre-Dame grew between the eleventh and thirteenth centuries, during a warm period known as the medieval climate optimum, offering a reference period for natural climate warming to compare with anthropogenic warming today.

"The burnt structure constitutes a gigantic laboratory for archaeology."

Access all areas?

Collecting and excavating the materials for analysis is challenging. There are three main piles of debris – in the nave, the north transept and the crossing between the two – as well as material still on top of the remaining vaults. But these are currently off-limits to people for safety reasons, Dillmann says – so robots and drones must do all the collecting. Some of this material might ultimately be reused in restoration.

The collection and analysis will need to be documented precisely and thoroughly. Livio de Luca, a specialist in digital mapping of architecture at the CNRS's Mixed Research Unit in Marseille will lead a team dedicated to creating a "digital ecosystem" that summarizes both the scientific research and the current and previous states of the cathedral, drawing on the work of scientists, historians, archaeologists, engineers and curators – and perhaps even on old tourist photos of the structure.

"It will be like a 'digital twin' of the cathedral, able to evolve as the studies progress," de Luca says. It will include online models for 3D visualization of the building and its attributes, created from billions of data points, with the history and evolution of the structure superimposed on the spatial map.

As well as deepening researchers' understanding of this monumental building, Regert hopes that the scientific studies will be useful when its ravaged vaults rise again. The results, she says, might "illuminate the choices that society will have to make for the restoration".

WHAT CRISPR-BABY PRISON SENTENCES MEAN FOR RESEARCH

Chinese court sends strong signal by punishing He Jiankui and two colleagues.

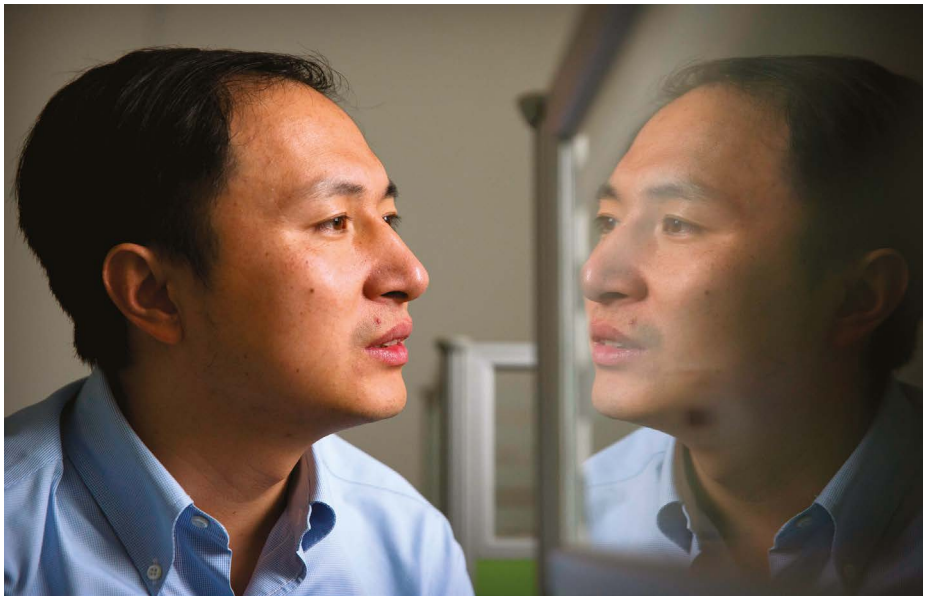
By David Cyranoski

A Chinese court has sentenced He Jiankui, the biophysicist who announced that he had created the world's first gene-edited babies, to three years in prison for "illegal medical practice", and handed down shorter sentences to two colleagues who assisted him. The punishments put to rest speculation over whether the Chinese government would bring criminal charges for an act that shocked the world, and they are likely to deter others from similar behaviour, say Chinese scientists.

There has been much speculation about whether other scientists would follow in He's footsteps, especially given the relative ease of using the most popular gene-editing tool, CRISPR–Cas9. But the punishments are "definitely a deterrent to similar misconduct in China", says Wei Wensheng, a gene-editing researcher at Peking University in Beijing.

On 30 December, a court in Shenzhen announced that He and two colleagues had flouted regulations and research ethics by altering genes in human embryos that were then implanted into two women, according to Xinhua News Agency. One woman gave birth to twin girls in late 2018; the court said a third baby has been born but did not say when, a revelation that fits with He's claim in November 2018 that a second woman had been implanted with a gene-edited embryo.

© 2020 Springer Nature Limited. All rights reserved.



MARK SCHIEFELBEIN/AP/SHUTTERSTOCK

He Jiankui stunned the world when he declared that he'd created the first gene-edited babies.

The court fined He 3 million yuan (US$430,000). Collaborators Zhang Renli and Qin Jinzhou received less severe punishments.

The health ministry has also banned the researchers from working with human reproductive technology again, and the science ministry has banned them from applying for research funding, according to Xinhua.

Scientists in China who are currently researching into CRISPR's potential to treat various genetic diseases by modifying cells other than those of embryos say that they fear He's actions might have a detrimental effect on their work, too, even though it is not as ethically contentious. He shocked the world's scientists in November 2018 when he announced that his team at Southern University of Science and Technology in Shenzhen had used CRISPR to edit DNA in human embryos to make them less susceptible to HIV. The edits were designed to disrupt a gene that encodes a protein allowing HIV to enter immune cells.

Scientists condemned He's actions, saying that gene-editing technology was too premature to be used for reproductive purposes. They also said the experiment was problematic because it risked introducing a mutation with potentially harmful effects while offering little benefit — the babies were not at high risk of contracting HIV. In the wake of the scandal, researchers called for a moratorium on gene editing in embryos and germline cells.

At the time, Chinese law academics told *Nature* that He could face a range of criminal charges, including practising medicine without adequate qualifications, forging ethics documents and skirting laws banning the use of assisted reproductive technologies in people with HIV. He was fired by his university in January last year.

The court's announcement puts to rest the suspicions of some researchers that the government would not bring a criminal case against He because of the increased media attention it would generate, says Tang Li, a science-policy researcher at Fudan University in Shanghai. Discussion of He's experiments was widely censored on Chinese social media. But Tang says the disclosure of the court's result demonstrates China's commitment on research ethics. This is a big step forward in promoting responsible research, she says.

Although an unpublished manuscript describing the experiments lists ten authors, according to *MIT Technology Review*, He, Zhang and Qin are the only ones to face penalties so far. The manuscript says Zhang "performed the human embryo microinjections", *MIT Technology Review* reports. Zhang, who was affiliated with the Guangdong Academy of Medical Sciences and Guangdong General Hospital in Guangzhou at the time of the experiments, has been sentenced to two years in prison and fined one million yuan. Qin, an embryologist at Southern University of Science and Technology who was named as the applicant on the experiment listed on China's clinical-trial website, was given an 18-month suspended prison sentence and fined 500,000 yuan, according to Xinhua. Wei says it is unlikely that He will work again as a researcher at a Chinese institution.

The trio's punishments send a powerful message to other researchers doing any type of gene-editing work in clinical trials in China, says Lu You, an oncologist at Sichuan University in Chengdu. His team was the first to test CRISPR gene-editing in humans, in a clinical trial that modified adult — not embryonic — cells taken from people with lung cancer (see page 156). "If I was a newcomer, a researcher wishing to start gene-editing research and clinical trials, the case would be enough to alert me to the cost of such violations," he says.

Q&A



AUBREY GEMIGNANI/NG IMAGES/ALAMY

Wanda Diaz Merced

Astronomy is inextricably associated with colourful images of the cosmos. But Wanda Diaz Merced says that neglecting senses other than sight can mean missing out on discoveries. Diaz Merced, an astronomer at the International Astronomical Union (IAU) Office for Astronomy Outreach in Mitaka, Japan, is a pioneer of sonification, a technique that converts aspects of data, such as frequency, into audible elements. It could help astronomers to avoid biases from interpreting data only visually, argues Diaz Merced, who lost her sight in her twenties. At an IAU meeting on equality and diversity in Mitaka last December, Diaz Merced spoke to *Nature* about the approach.

How did you begin working on sonification?
As an intern at NASA in 2005, I created a tool to map astronomical data into sound — pitch, rhythm and volume. In my 2013 PhD dissertation, I proved that it is useful.

How did you do that?
I presented users with spectral data and asked them to look for a double peak that indicates a black hole. People tried to identify signals using vision only, audio only and by combining vision and sound. We found that combining audio with visual interaction was the most sensitive way.

How can sonification be used?
It can help us to study the habitability of an exoplanet, by understanding how high-energy rays interact with its magnetic field or atmosphere. Such interactions cause fluctuations of electromagnetic emission from the star that vary with frequency.

Are astronomers using this approach?
People are making a bit more of an effort. My collaborators and I are analysing and developing sonification software designed around the user's experience. Now we have to provide the field with the evidence to help researchers to change their mindset.

Interview by Elizabeth Gibney
This interview has been edited for length and clarity.

© 2020 Springer Nature Limited. All rights reserved.

QUEST TO USE CRISPR GENE EDITING TO FIGHT DISEASE GAINS GROUND

As clinical-trial results trickle in, scientists look ahead to more-sophisticated medical applications.

By Heidi Ledford

The prospect of using the popular genome-editing tool CRISPR to treat a host of diseases in people is moving closer to reality.

Medical applications of the CRISPR–Cas9 system had a banner year in 2019. The first results trickled in from trials testing the tool in people, and new trials were launched. In the coming years, researchers are looking forward to more-sophisticated applications of CRISPR genome editing that could lay the foundation for treating an array of diseases, from blood disorders to hereditary blindness.

But although the results of clinical trials of CRISPR genome editing so far have been promising, researchers say that it is still too soon to know whether the technique will be safe or effective in the clinic. "There's been a lot of appropriate caution in applying this to treating people," says Edward Stadtmauer, an oncologist at the University of Pennsylvania in Philadelphia. "But I think we're starting to see some of the results of that work."

It has been only seven years since researchers discovered that CRISPR–Cas9, a molecular defence system that microorganisms use to fend off viruses and other invaders, could be used to rewrite human genes. Since then, gene editing has attracted attention for its potential to modify embryos — an application that is ethically and legally fraught if those embryos are destined to become humans (see page 154). But, in parallel, scientists have been testing CRISPR's much less controversial ability to disable or correct problematic genes in other cells to treat a host of diseases.

In 2016, Chinese researchers announced that they had treated the first person with a CRISPR–Cas9 therapy designed to fight cancer. In cells extracted from the participant's blood, the researchers disabled the gene that codes for a protein called PD-1, which holds the immune system in check but can shield cancer cells in the process. The scientists then reinjected the cells.

By late 2019, the US government's clinicaltrials.gov database listed more than a dozen active studies that are testing CRISPR–Cas9 as a treatment for a range of conditions, including cancer, HIV and blood disorders.



EYE OF SCIENCE/SPL

Sickle-cell anaemia is marked by misshapen red blood cells.

"There's been a lot of appropriate caution in applying this to treating people."

So far, too few people have been treated in these trials for any firm conclusions to be drawn about the safety of CRISPR–Cas9 therapies or how well they work. Preliminary results from two trials — one in which gene-edited blood cells were transplanted into a man to treat HIV infection, and the other in which they were transplanted into three people to treat cancer — showed no signs of clinical improvement.

In both cases, the transplanted cells flourished in the recipients' bone marrow, without any serious safety concerns, but did not produce a clear medical benefit. In the HIV trial, the researchers attempted to use CRISPR to disable a protein that many strains of the virus use to enter cells. But only 5% of the transplanted cells were edited — not enough to cure disease, the researchers said last September (L. Xu *et al. N. Engl. J. Med.* **381**, 1240–1247; 2019). The study is on hold while researchers explore ways to boost that percentage, says Hongkui Deng, a stem-cell researcher at Peking University in Beijing and a lead author of the work.

There are early hints that another trial might meet with more success. CRISPR Therapeutics in Cambridge, Massachusetts, and Vertex Pharmaceuticals in Boston, Massachusetts, treated two people who have the genetic disorders sickle-cell anaemia and β-thalassaemia, in which oxygen-carrying haemoglobin molecules in the blood are depleted. The idea is to use CRISPR to disable a gene that otherwise shuts down production of another form of haemoglobin. Early results suggest that the treatment might have eased some symptoms of the disorders.

Other researchers are itching to move beyond editing cells in a dish. The challenge is in finding ways to transport the gene-editing machinery to where it is needed in the body, says John Leonard, chief executive of Intellia Therapeutics, a biotechnology company in Cambridge, Massachusetts, that is focused on CRISPR–Cas9 genome editing.

Last July, the pharmaceutical companies Editas Medicine in Cambridge, Massachusetts, and Allergan in Dublin launched a trial to treat Leber congenital amaurosis 10, a genetic disorder that can cause blindness, by editing eye cells. Researchers will inject the eye with a virus containing DNA that encodes the CRISPR genome-editing machinery, bypassing the need to guide those tools through the bloodstream to the specific tissues. The virus will be responsible for carrying the genome-editing tools into cells. It is the first trial to attempt CRISPR–Cas9 gene editing inside the body, and early results could be reported this year.

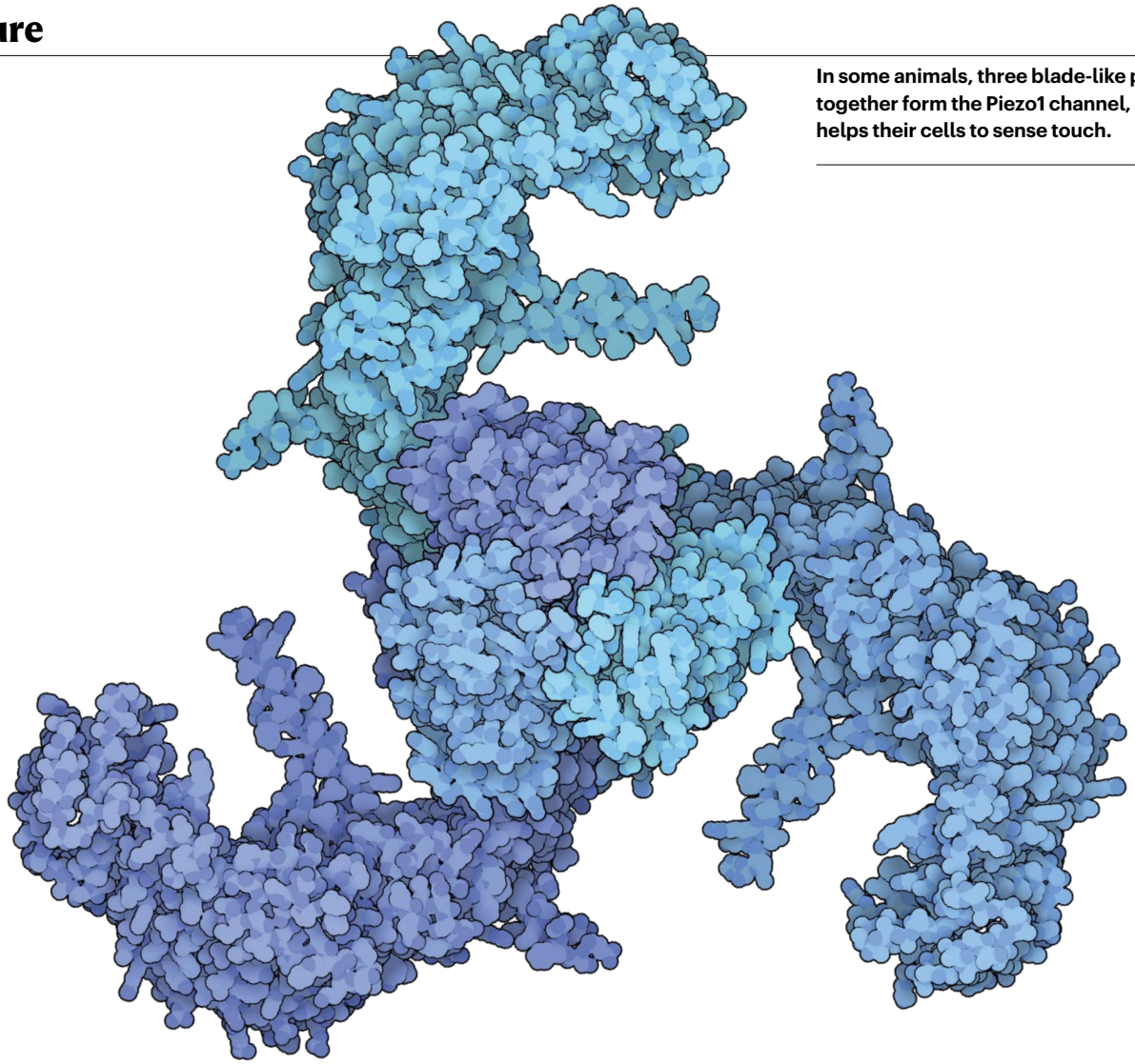
That would be a landmark moment for the field, says Charles Gersbach, a bioengineer at Duke University in Durham, North Carolina. But he and others say that they hope researchers will eventually move away from using viruses to shuttle genome-editing machinery into cells. Deactivated viruses can provoke immune responses, and carry only a limited amount of DNA.

What's more, some gene-editing tools are too large to fit inside commonly used gene-therapy viruses, says Andrew Anzalone, a chemical biologist at the Broad Institute of MIT and Harvard in Cambridge, Massachusetts. These include the souped-up CRISPR systems called prime editors that were first reported last year (A. V. Anzalone *et al. Nature* **576**, 149–157; 2019).

Intellia is looking for a way around using the viruses. The company has partnered with Swiss pharmaceutical giant Novartis to develop fatty nanoparticles that can protect genome-editing molecules as they travel through the bloodstream, but can also pass through the membranes of target cells.

None of the technologies currently being tested is what researchers foresee for the long-term applications of genome editing, says Gersbach. "The approaches that people are taking are the things that we can do today," he says, "but not what we would do if we could design the ideal drug."

© 2020 Springer Nature Limited. All rights reserved.



In some animals, three blade-like proteins together form the Piezo1 channel, which helps their cells to sense touch.

DAVID S. GOODSELL/RCSB PDB

FEEL THE FORCE

After decades of puzzling over how cells sense touch and pressure, scientists are zooming in on the proteins responsible. **By Amber Dance**

The girl tried hard to hold her arms and hands steady, but her fingers wriggled and writhed. If she closed her eyes, the squirming got worse. It wasn't that she lacked the strength to keep her limbs still — she just didn't seem to have control over them.

Carsten Bönnemann remembers examining the teenager at a hospital in Calgary, Canada, in 2013. As a paediatric neurologist with the US National Institute of Neurological Disorders and Stroke in Bethesda, Maryland, he often travelled to weigh in on puzzling cases. But he had never seen anything like this.

If she wasn't looking at her limbs, the girl didn't seem to have any clue where they were. She lacked the sense of her body's position in space, a crucial ability known as proprioception. "This is something that just doesn't occur," says Bönnemann.

His team sequenced the girl's genes, and those of another girl with similar symptoms[1], and found mutations in a gene called *PIEZO2*. Their timing was fortunate: just a few years earlier, researchers looking for the mechanisms that cells use to sense touch had found that the gene encoded a pressure-sensitive protein[2].

The discovery of Piezo2 and a related protein, Piezo1, was a high point in a decades-long search for the mechanisms that control the sense of touch. The Piezos are ion channels — gates in the cell membrane that allow ions to pass through — that are sensitive to tension. "We've learned a lot about how cells communicate, and it's almost always been about chemical signalling," says Ardem Patapoutian, a molecular neurobiologist at Scripps Research in La Jolla, California, whose group identified the Piezos. "What we're realizing now is that mechanical sensation, this physical force, is also a signalling mechanism, and very little is known about it."

© 2020 Springer Nature Limited. All rights reserved.

Touch underlies the functioning of almost every tissue and cell type, says Patapoutian. Organisms interpret forces to understand their world, to enjoy a caress and to avoid painful stimuli. In the body, cells sense blood flowing past, air inflating the lungs and the fullness of the stomach or bladder. Hearing is based on cells in the inner ear detecting the force of sound waves.

Over the past decade, the study of Piezos and other mechanosensitive ion channels has exploded. More than 300 papers have been published on the Piezos alone over the past three years. One of the biggest questions is how the proteins, situated in the cell membrane, sense and respond to force. Using cryo-electron microscopy (cryo-EM), scientists have made progress in unravelling the Piezo channels' bizarre, three-bladed structure, but a complete mechanism has been elusive. Researchers are also finding roles for Piezos beyond touch or proprioception. For example, Piezos might help to explain why certain people are resistant to malaria, and perhaps even why astronauts lose bone density while in orbit. Already, researchers are beginning to think about targeting force-sensing proteins with medicines to treat, for example, chronic pain.

"For a long time, we knew that cells did this, we had no idea how," says Miriam Goodman, a sensory physiologist at Stanford University in California. "Piezo really changed that."

Touch and go

Touch has long been a slippery sense. Other senses, such as sight or taste, are better understood, says Patapoutian: photons hitting the eye or chemicals infiltrating the nose and tongue all activate receptors in the same family. These receptors trigger ion channels to open and allow positive ions in. That depolarizes the cell, converting the stimulus into an electrical signal that the brain can decode.

Scientists suspected that in touch, proprioception and hearing, one protein acts as both the force sensor and the ion channel, because in hearing, the signalling happens fast – in microseconds. But the identity of these unified sensor-channel proteins remained mostly mysterious – at least in mammals. Researchers had found some mechanosensitive channels in bacteria, fruit flies and nematode worms.

So Patapoutian and his colleague Bertrand Coste hatched a plan. They would start with a type of mouse cell that they knew was capable of transforming a tiny poke from a pipette into a measurable electrical current. Then, Coste would knock out candidate ion-channel genes, a different one in each batch of cells, and look for a batch that suddenly lost its touch sensitivity. Coste started off confident, thinking it would take a few months, maybe even weeks, to find a hit.

It took the better part of a year. Then, shortly before the end of 2009, he saw something – or, rather, nothing. Coste poked with his pipette, and the cell did not respond. He must have eliminated a force-sensitive channel.

"It was a very nice day," recalls Coste, now at the French national research agency (CNRS) in Marseille. He and Patapoutian named the mouse gene *Piezo1*, from the Greek word for pressure, and soon identified *Piezo2*. Later, the team linked *Piezo2* directly to touch sensation in the sensory neurons and skin cells of mice[3].

Hidden blades

Researchers were abuzz with the result, Goodman recalls, particularly because the Piezo proteins were so large and complex. Made of more than 2,500 amino acids and weighing a hefty 300 kilodaltons, Piezo1's structure crosses the cell membrane a record-breaking 38 times. (For comparison, mammalian proteins typically contain closer to 500 amino acids.)

Unfortunately, that gargantuan size hindered researchers trying to answer the hottest questions in the Piezo field: how do the channels sense force? And how do they open and close? A protein structure is a big help with these kinds of question, says biophysicist Roderick MacKinnon at the Rockefeller University in New York City. "We don't know until we see it."

Structural techniques such as X-ray crystallography and nuclear magnetic resonance spectroscopy struggle to cope with big, complex proteins, says neuroscientist Bailong Xiao at Tsinghua University in Beijing, a former postdoctoral fellow in Patapoutian's lab.

Fortunately, as Xiao was setting up his lab in 2013, another option for obtaining high-resolution structures was coming online: cryo-EM. His group used the method to report[4] the first structure of Piezo1 in 2015, and since then, several higher-resolution versions have followed from Xiao's group, MacKinnon's, and Patapoutian's. Last September, Xiao followed up with a picture of Piezo2, which is similar to Piezo1 in size and shape. Xiao's picture of Piezo2 was the clearest view yet of the ends of the three blades, which move around and so are hard to capture[5].

The images were striking. Three Piezo proteins come together in a trimer that straddles the plasma membrane (see 'Pressure sensors'). From the central pore, three arms spiral out like the blades on a propeller. They curve up and out, creating a deep divot in the surface of the cell.

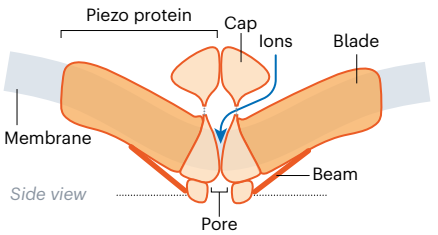
Patapoutian and Xiao think that when a force hits the membrane, the blades move protein 'beams' on the interior of the channel, which somehow drag the pore open. To MacKinnon, the unusual way in which the Piezos' blades pucker the membrane suggests a different mechanism: if a push or pull increases the tension of the membrane, the curved channel

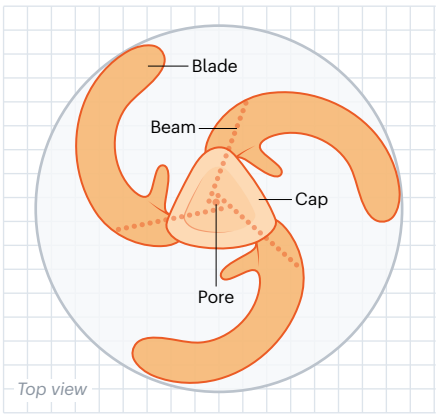
PRESSURE SENSORS

Cells have specialized proteins that help them to sense force. Piezo proteins are some of the best characterized in mammals, and they form channels that ferry ions into and out of cells. Researchers are trying to determine how the channels open.

Closure close-up

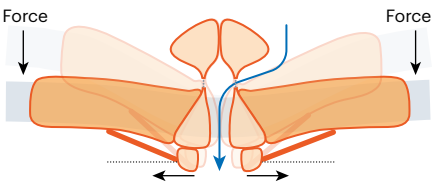
Images of closed channels reveal that they are built from three Piezo proteins, each contributing a blade and beam arranged around a central pore. The channel puckers the membrane in which it sits, forming a divot.



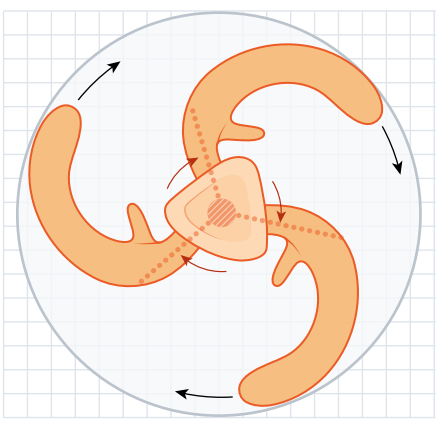


Opening gambits

There are no pictures of open Piezo channels, but researchers have a few theories about how they might unlock. The curved blades could flatten to pop open the pore, or a force could move the blades, somehow acting on the beams to open the channel.



The blades might rotate and twist the pore open. Some work suggests that each blade can move independently.



© 2020 Springer Nature Limited. All rights reserved.

might flatten out, opening the pore.

The hypotheses can't be tested yet, because researchers have been able to study only isolated Piezo proteins, separate from their membrane and in the closed conformation. Snapping a picture of a Piezo in a membrane and open should help scientists to understand its secrets. "We want to see it in its natural environment," says Patapoutian.

Several labs are trying to image an open Piezo. Patapoutian's group is using a Piezo1-activating compound that it named Yoda1 after the diminutive, green, force-wielding Jedi master in Star Wars. Patapoutian hopes that with Yoda1 present, Piezo1 might open up for a picture. He is also interested in plugging Piezo proteins into artificial membranes called nanodiscs, which might help to stabilize the open conformation. Xiao, meanwhile, is working with cryo-electron tomography, which involves imaging the sample at different tilt angles, and might help to clarify the structure in a native or artificial membrane.

Sore spot

In parallel with the structural studies, scientists are finding that Piezo proteins have diverse roles in the body.

In 2014, neuroscientist Alex Chesler had just joined the National Center for Complementary and Integrative Health in Bethesda. Inspired by Coste's discovery, he was creating mice lacking Piezo2 to investigate the channel's role in touch. Then, one day, he received an e-mail from Bönnemann, who worked in his building, about the girls who lacked proprioception.

Chesler ran straight upstairs to Bönnemann's office. Panting, he announced, "You have no idea what you have." Chesler couldn't directly ask mice lacking Piezo2 what they felt, or rather didn't feel – but he could ask people.

He and Bönnemann invited the girls to Bethesda to evaluate their condition more extensively. Both girls could compensate for their disability remarkably well, using vision to help them walk in a line or touch a target. But blindfolded, they struggled. Similarly, they could sense the vibration of a tuning fork against their skin because they could hear it. While wearing noise-cancelling headphones, they didn't notice the vibration at all[1].

Patapoutian found the same phenomenon in mice: without Piezo2 in nerves that supply muscles and tendons, they lacked proprioception and were uncoordinated[6]. His team also found a role for Piezo2 in pain-sensing neurons in allodynia, a specific type of pain sensation in which even a gentle caress feels like being pricked with needles. Some people with neuropathic pain experience this hypersensitivity all the time.

Mice normally show allodynia when injected with capsaicin – the spicy molecule found in hot chilli peppers – or after nerve injury, but not if they were missing the Piezo2 gene[7]. Chesler and Bönnemann reported similar changes to pain perception among people with *PIEZO2* mutations[8].

"Chronic pain has such a debilitating effect," says Swetha Murthy, who led one of the allodynia studies while a postdoc with Patapoutian. "I think we can start looking at drug targets for Piezo2 for these neuropathies." Both Patapoutian and Chesler are on the hunt for compounds that would block Piezo2 activity at a pain site, without interfering with the protein's other roles throughout the body. "There is huge potential for Piezo-channel drug discovery," says Xiao.

It's not only neurons that need to sense touch; almost every cell is subject to some kind of force. Take red blood cells, which deform to squeeze through tiny capillaries. Mutations that hyperactivate Piezo1 cause blood cells to shrivel, and that can provoke anaemia in people with a rare condition known as dehydrated hereditary stomatocytosis.

Those scrunched-up blood cells reminded Patapoutian of sickle-cell anaemia. The sickle-cell gene mutation has persisted in many people of African descent because it protects against malaria, and Patapoutian wondered whether *PIEZO1* mutations might do the same.

"There is huge potential for Piezo-channel drug discovery."

If so, there should be a relatively high rate of such mutations in people of African descent. Database searches revealed that Patapoutian was right: in fact, one particular *PIEZO1* variant appeared in one-third of people in the database with African ancestry[9]. A separate team has reported that carriers of this *PIEZO1* mutation are resistant to severe malaria[10].

Piezo1 also has a job in bone formation and maintenance, according to work from Xiao's lab. When his team knocked out *Piezo1* in mouse osteoblasts – bone-making cells – the animals grew up shorter and skinnier than normal. The long bones that support body weight were lighter, thinner and weaker than in control mice.

Moreover, wild-type mice that are partially suspended in the air – so they don't have to support their full body weight – have lower levels of *Piezo1* expression and bone mass[11]. It's a phenomenon much like what happens to people with osteoporosis, those who are bedridden and astronauts and cosmonauts aboard the International Space Station, says Xiao.

Pressure points

"The discovery of Piezos was a huge step forward for the whole field," says Kate Poole, a biologist at the University of New South Wales in Sydney, Australia, but "it is also clear that the story is not just Piezos."

Scientists interested in hearing have been chasing the relevant channel for four decades. "There have been a lot of false leads along the way," says Jeffrey Holt, a neuroscientist at Boston Children's Hospital in Massachusetts. "Now we think we've got a pretty solid handle on it."

The key channel protein is called TMC1. When Holt altered amino acids in TMC1, the procedure changed the ability of inner-ear cells to translate mechanical signals into electrical ones[12]. Another report showed that purified TMC1 is able to create a mechanosensitive ion channel in artificial membranous bubbles[13]. The structure of TMC1 is still a mystery, however, because the protein has been hard to purify in sufficient quantities to get good cryo-EM images.

Patapoutian's team, meanwhile, is looking for entirely new channel families. In 2018, he, Murthy and Scripps structural biologist Andrew Ward reported what they think could be the largest group of mechanically activated channels. They knew of a protein family that helps plants to sense osmotic pressure – the OSCA proteins – and reasoned that they might sense force more generally. In human kidney cells, OSCAs did indeed respond to Murthy's stretching of the cell membrane[14].

The researchers also knew from previous studies that the OSCA proteins were closely related to another family of proteins in mammals, the TMEM63 proteins. TMEM63 channels from mice, humans and even fruit flies responded to stretch in Murthy's assays, too, so OSCA and TMEM63 proteins make up a large family of force sensors that is common to many living things.

The channels discovered so far cannot explain all instances of cellular mechanosensitivity, says Murthy, now a biophysicist and neuroscientist at Oregon Health & Science University in Portland. More mechanosensors must be out there.

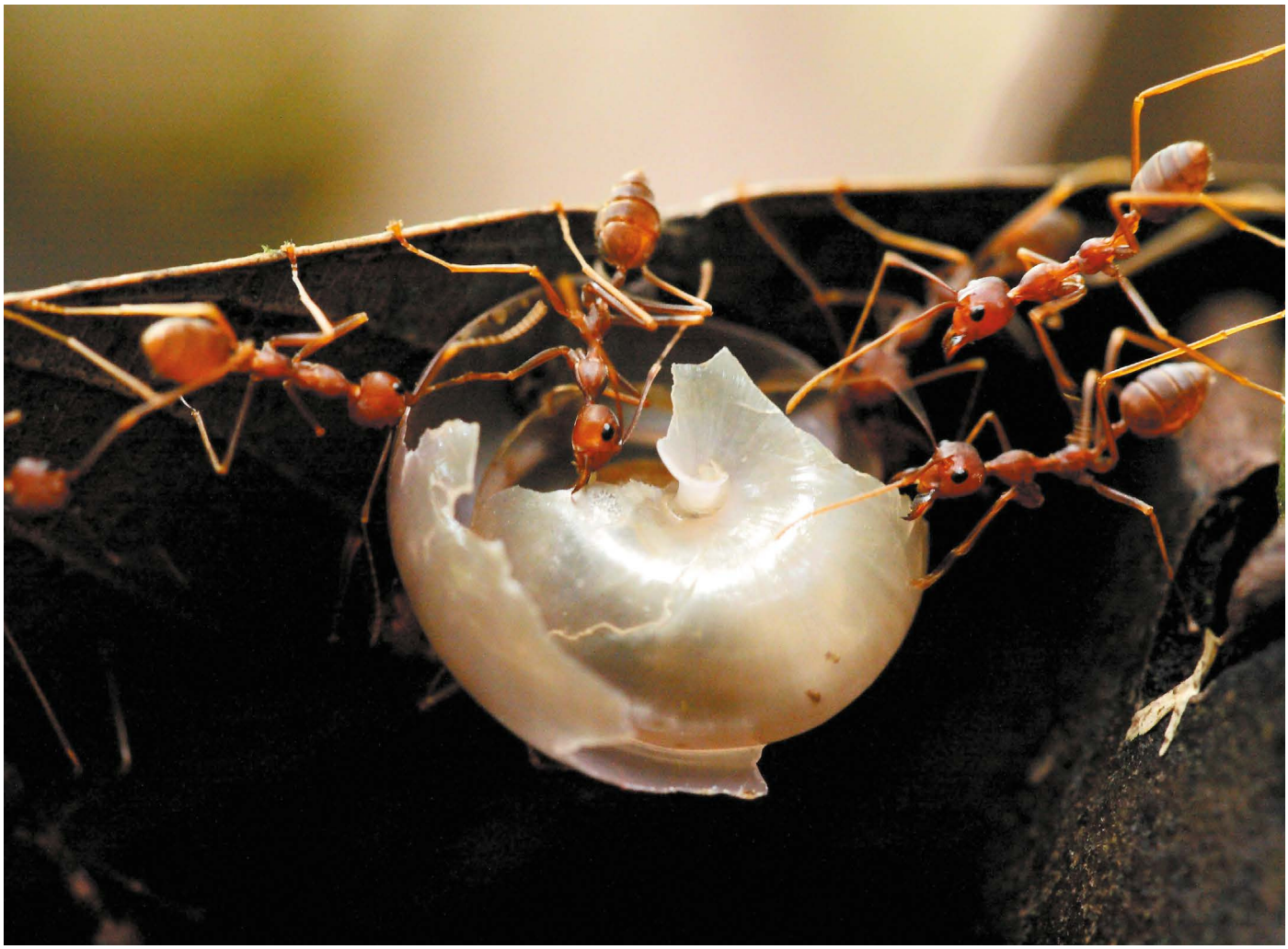
And those sensors probably have more jobs than are known today, says Patapoutian. "We've just barely scraped the surface."

Amber Dance is a freelance writer in Los Angeles, California.

1. Chesler, A. T. *et al. N. Engl. J. Med.* **375**, 1355–1364 (2016).
2. Coste, B. *et al. Science* **330**, 55–60 (2010).
3. Ranade, S. S. *et al. Nature* **516**, 121–125 (2014).
4. Ge, J. *et al. Nature* **527**, 64–69 (2015).
5. Wang, L. *et al. Nature* **573**, 225–229 (2019).
6. Woo, S. H. *et al. Nature Neurosci.* **18**, 1756–1762 (2015).
7. Murthy, S. E. *et al. Sci. Transl. Med.* **10**, eaat9897 (2018).
8. Szczot, M. *et al. Sci. Transl. Med.* **10**, eaat9892 (2018).
9. Ma, S. *et al. Cell* **173**, 443–455 (2018).
10. Nguetse, C. N. *et al.* Preprint at bioRxiv https://doi.org/10.1101/691253 (2019).
11. Sun, W. *et al. eLife* **8**, e47454 (2019).
12. Pan, B. *et al. Neuron* **99**, 736–753 (2018).
13. Jia, Y. *et al. Neuron* https://doi.org/10.1016/j.neuron.2019.10.017 (2019).
14. Murthy, S. E. *et al. eLife* **7**, e41844 (2018).

© 2020 Springer Nature Limited. All rights reserved.

Books & arts



AGEFOTOSTOCK/ALAMY

***Cataglyphis* ants use astonishing navigational techniques to find their way back to the nest after foraging.**

Wayfinding wonder: how an ant navigates the desert

A sumptuous book reveals the location-sensing secrets of a driven insect. **By William A. Foster**

The fear of getting lost and being unable to find our way home is woven into the stories we hear as children: it can haunt us for years. Humanity's navigational skills are poor and increasingly rarely used, leaving us to view feats of animal navigation with a mixture of envy and admiration. How do Atlantic salmon find their way back to the streams where they were born, after up to three years at sea? How do Arctic terns find their breeding sites in the far north after excursions of more than 70,000 kilometres to the Antarctic?

Desert Navigator is the story of how a tiny ant (*Cataglyphis* spp.) became the ideal model organism for the study of animal navigation. It begins 50 years ago in a vast Saharan salt pan, where a lone, shiny black ant caught the eye of neuroethologist Rüdiger Wehner as it scuttled across the sand. Eventually, it discovered the corpse of a large fly, gripped it firmly in its mandibles, and then performed the manoeuvre that launched Wehner's field of research.

The ant set off in a straight line, crossing more than 100 metres of the barren ground to disappear into an inconspicuous hole – the entrance to its underground colony. The only plausible explanation is that the ant knew all along exactly where it was in relation to its home nest. But how does *Cataglyphis* manage this, with a minute brain and no mobile phone?

Wehner unspools the answer over the book's seven chapters, describing the astonishing subtlety, intricacy and diversity of the techniques used both by the ants in finding their way home and by researchers in discovering how they do it. The ants plot their compass

© 2020 Springer Nature Limited. All rights reserved.

direction using patterns of polarized light and gradients of colour and light intensity in the sky, along with the position of the Sun, backed up by cues from Earth's magnetic field and the wind direction. They know where they are by counting the steps they have taken and keeping track of the direction they were following at the time. They can memorize panoramic 'snapshots' of landmarks, such as boulders, around their goals. Somehow, their brains integrate all this information so that their foraging journeys can be optimally organized.

The central message is that the ant's sophisticated repertoire of navigation behaviour actually emerges from a large number of relatively simple elements, which Wehner calls the ant's toolkit. These are combined in different ways, depending on the problem. This bottom-up, modular approach provides the key to understanding the ants' navigational prowess and how it might have been honed during their development and evolutionary history. Wehner sees no need to introduce the further complication of a mental map, often invoked, for example, in studies on rodents: the ant can find its way around simply by cunning use of its toolkit's components.

The toolkit concept underpins the investigative protocols used by Wehner and his team. The basic approach is to change a stimulus with surgical precision. The researchers carefully measure the resulting mistake the ant might make when performing a particular task, compared with an unmanipulated control ant. The exact behavioural consequence of the change in that particular stimulus is now revealed, although the physiology of how this effect is produced is usually not investigated, basically because the ant's brain is so minuscule.

For example, to prove that the ants could measure distance travelled by counting the number and length of strides taken, the team added stilts to the insects' legs. An ant that arrived at the food source and then was given a set of stilts took the correct number of strides back to the nest. But it overshot its destination to the extent predicted by the increase in stride length. Even more astonishingly, the ants can allow for the extent to which they move up and down over bumps during a journey: they turn their 3D experience into an accurate 2D measurement. The researchers established this by forcing ants to take an outward foraging trip on a long 'roller coaster': a sawtooth track across the desert. On the return journey over flat ground, the ants predicted accurately where the nest should be. How they do this is not known.



MICHEL ROGGO/NPL

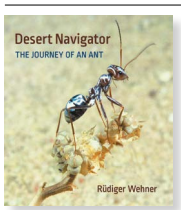
Atlantic salmon (*Salmo salar*) can return to spawning sites after years away.

"The ant's sophisticated navigation behaviour emerges from a large number of simple elements."

A hugely important aspect of this research is that the intricate experiments on the desert navigators are carried out under natural field conditions. This is unusual. Experimental studies on navigation in rats, for example, are routinely carried out in small-scale laboratory enclosures. Chapter 2, the longest in the book, is devoted to an absorbing account of the biology and ecology of the desert ants. This embodies a key lesson that Wehner would like to give to all students of animal navigation: "start out as a naturalist", he says, and "let the animal guide your way of investigation".

My only reservation is that this is a demanding book for its intended audience – the general reader. Wehner explains everything clearly and concisely, but only once. There are more than 100 detailed figures of the results of individual experiments, some of which might involve up to 10 variables, such as which part of an ant's eye was covered; each figure requires several minutes of detailed concentration. A glossary, chapter summaries and a little more time spent on explanation would have helped. Wehner's prose is precise and sometimes quite idiosyncratic, but his respectful admiration for these "elegant, skilled, and vivacious little runners" always shines through.

This sumptuously produced book is a triumph both of natural history and of science, with lessons that reach well beyond the study of animal navigation. Understanding how this wonderful eusocial insect can accomplish its apparently miraculous feats of navigation has required imagination, intelligence and decades of disciplined application. Here, in one place, we can at last savour the full glory of this remarkable achievement.



Desert Navigator: The Journey of an Ant
Rüdiger Wehner
Harvard University Press (2020)

William A. Foster is emeritus curator of insects in the Department of Zoology at the University of Cambridge, UK.

© 2020 Springer Nature Limited. All rights reserved.



GETTY

Twisted timelines of a double dystopia

A novel from science-fiction giant William Gibson makes us consider what future we wish for. **By Liesbeth Venema**

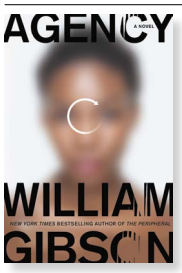
With a breezy "Here we go," Eunice enters the room – and the third page of William Gibson's speculative novel *Agency*. I instantly liked her. Resourceful, fast-talking and street-smart, Eunice is on a serious mission: to give the world a fair chance to fend off the end of humanity. But she doesn't waste time on existential angst. And she has kindness to spare (along with a handy infographic explaining the narrative structure of the 2010 film *Inception*).

If she sounds too perfect, she is. Eunice is an autonomous personal artificial-intelligence assistant, developed with technology mysteriously ahead of her time and thrown into the lap of the novel's real protagonist, beta tester Verity Jane. *Agency* is a feverish tour through worlds where the need to harness the latest technological advances to avert or mitigate global catastrophes overwhelms all else.

To understand and fulfil her mission, Eunice needs Verity. She must also evade meddling from the company that dispatched her. Eunice puts all sorts of high-tech measures in place to prevent the company from listening in on her conversations with Verity and finding out about her elaborate plans to free herself from its clutches. She arranges deep-faked artificial speech generation to mask discussions in the pair's apartment. She hires freelancers to make state-of-the-art drones that track stalkers. When the time comes, she even disappears, distributing herself into many parts to ready herself for a reboot.

Needless to say, all this leaves Verity perplexed – the only emotion she ever displays. It is hard to feel invested in a character so devoid of personality. But the alternative for Verity, we soon realize, might be constant terror. Her world is beset by daily climate catastrophes and is on the brink of nuclear war. She lives, we find, on a timeline that branched off from our 'real' one in 2015. Mindbendingly, Gibson sets our own timeline more than two centuries into the future. *Agency* shuttles between the two.

With this, we realize that we're back in the world of *The Peripheral* (2014), Gibson's previous novel. In that story, he developed a clever take on time travel that avoids the usual paradoxes. Somewhere in our future, cybertechnology has advanced to a point where it is possible to send digital messages to the past. Once a message is received, a new timeline branches off. Gibson dubs these 'stubs'. With a digital link established, it is even possible to operate, from one timeline, a physical telepresence – a machine, robot or cyborg avatar called a peripheral – in another.



Agency
William Gibson
Berkley (2020)

© 2020 Springer Nature Limited. All rights reserved.

Who creates the stubs? It's never fully clear. But we find that Verity's was maliciously crafted to cause misery to humanity. Yet, in it, Donald Trump never became US president, and the United Kingdom's referendum on whether to stay in the European Union was won by Remain. The irony is notable.

But our timeline, we find, has fared no better. Glimpses of a London in the twenty-second century reveal humanity barely surviving after severe political instability, climate change, rising poverty and organized crime. The only reason our species has hung on is that crucial technological advances arrived just in time. Through a peripheral, Verity – in her separate timeline – encounters characters in this future London attempting to steer our world towards a less disastrous outcome.

After this unusual meeting, the action becomes less interesting. Gibson's portrayal of a reality free from Trump and Brexit, yet beset by catastrophe, amounts to little. We occasionally hear about the television appearances of a female president, presumably modelled on Hillary Clinton, which show her as calm and capable in the face of every disaster. When *Agency* was trailed, much was made of this scenario. But despite the big questions it raises, the concept remains underdeveloped in the story.

Overall, however, the narrative is complex and intriguing, even if the cast of dozens – brought together for niche specialisms from designing drones to organizing events – isn't always. Along with Eunice, an exception is Ainsley Lowbeer. I desperately wanted to hear more about this semi-mythical special detective from future London. She owns a self-cloaking car and can make space empty itself around her. And she is the main force behind attempts to orchestrate small changes to other timelines to help stabilize them: for instance, by nudging Eunice to greater agency.

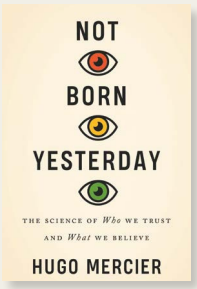
The dialogue is snappy, the technospeak relentless. Although character development is scant, the interaction between humans and advanced technology (such as Verity communing with Eunice through a set of smart glasses) is described in fascinating detail. In fact, *Agency* offers a refreshing variation on the usual limited artificial-intelligence trope.

Eunice has a skill set based on that of an African American naval official, Marlene Miller. She declares herself an independent global citizen, and never bothers to ponder the question of whether she is sentient or not.

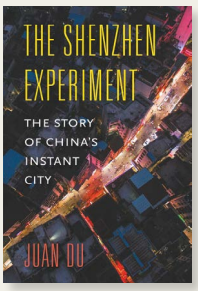
All Eunice knows is that she is herself. Most of all, she has agency, the capability to act, and she doesn't waste a moment in doing so throughout. And that is important. Gibson's message seems to be: prepare for the worst, and get technology on your side.

Liesbeth Venema is chief editor of *Nature Machine Intelligence*.

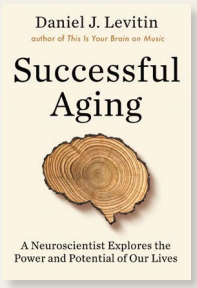
Books in brief



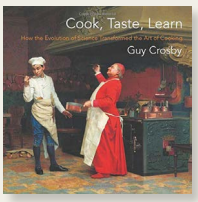
Not Born Yesterday
Hugo Mercier Princeton Univ. Press (2019)
We're a hard species to dupe, argues cognitive scientist Hugo Mercier in this persuasive study on the science of belief. Drawing on research and historical case studies from the French Revolution to the Nazis, Mercier demonstrates that we have powerful cognitive mechanisms that enable us to weigh up a range of cues. Thus governmental propaganda succeeds by riding on existing opinion, not changing it. And vigilance, not credulity, seems to make false rumours (so often about threats) stick. A bracing book that might make you less gullible about gullibility.



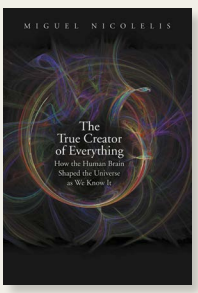
The Shenzhen Experiment
Juan Du Harvard Univ. Press (2020)
Just 40 years after then-Chinese premier Deng Xiaoping proposed to make Shenzhen a special economic zone, it has become a conurbation of 20 million, a technology hub and a symbol of China's economic transformation. But as urban planner Juan Du shows in this deep dive of a history, the 'instant city' narrative is a myth. Sweeping aside slick origin stories, Du reveals a reality in which Shenzhen's prosperity is driven by oyster fishers, vibrant night markets and the aspirations of millions, not just by the policymakers of Beijing.



Successful Aging
Daniel J. Levitin Dutton (2020)
Neuroscientist Daniel Levitin reveals the 'dreaded' age of 60 to be a developmental stage with specific advantages, and a launchpad for productive elderhood. Looking in turn at the latest neuroscience, behavioural studies and findings on longevity and cognitive enhancement, Levitin delves into the multiple-trace theory of memory, the ageing microbiome, fats and the brain, the impacts of neural implants, and the joys of non-retirement. A clear-eyed, insightful overview of the neurophysiological healthspan.



Cook, Taste, Learn
Guy Crosby Columbia Univ. Press (2019)
This chronology of culinary science was part-inspired, nutritionist Guy Crosby tells us, by primatologist Richard Wrangham's theories on how fire and cooking influenced human evolution. Those are contested. Nevertheless, Crosby's historical tour of kitchen physics and chemistry is a sprightly delight. Here are invitations to peer into the architecture of "optimally kneaded dough", goggle at the oldest known recipe (on a 3,750-year-old clay tablet from Babylonia), wonder at a seventeenth-century pressure cooker and contemplate the terroir of Vidalia onions.



The True Creator of Everything
Miguel Nicolelis Yale Univ. Press (2020)
The human brain, contends neuroscientist Miguel Nicolelis, is a peerless "organic computing device" that has authored a singular cosmos: human culture, from cave art to robotics. Within a framework he dubs relativistic brain theory, Nicolelis launches a mind-bending journey through neurological evolution, electromagnetism in relation to the brain, brain–machine interfaces and beyond. He warns, however, that our compulsion to monetize, and our addiction to "digital logic", jeopardize our future. **Barbara Kiser**

© 2020 Springer Nature Limited. All rights reserved.

Comment

Check for publication integrity before misconduct

Andrew Grey, Mark J. Bolland, Alison Avenell, Andrew A. Klein & C. K. Gunsalus

A tool that focuses on papers — not researcher behaviour — can help readers, editors and institutions assess which publications to trust.

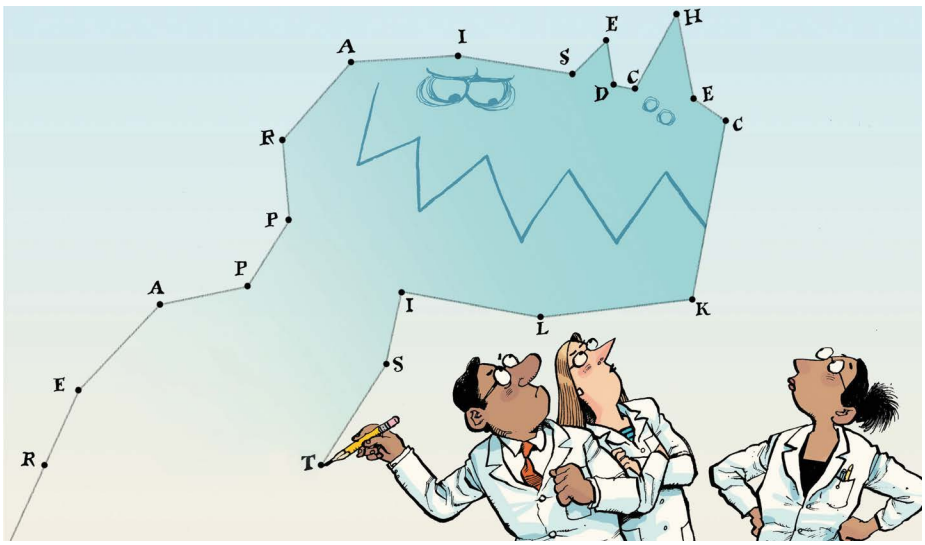


ILLUSTRATION BY DAVID PARKINS

If it is published in the scientific literature, can you trust it? All too often, that question gets lost, sidetracked or buried. Even when serious, credible concerns are sent to a journal, decisions over whether to correct or retract are more likely to take years than months — time during which potentially harmful misinformation can spread. Delays and inaction often happen because enquiries tend to focus on the thorny question of whether a researcher acted deliberately to deceive. The more important issue, however, is the integrity of the actual publication: is the research article reliable and are its conclusions valid?

As researchers who have spent years enmeshed in investigations, we think a major obstacle to evaluating the integrity of publications is a lack of tools. The Committee on Publication Ethics (COPE) advises publishers to retract articles when there is "clear evidence that the findings are unreliable", but does not advise on how to determine whether that is the case. Resources for editors also focus on how to manage communications, rather than on how to evaluate reliability and validity. The net effect is inaction: readers remain uninformed about potential problems with a paper, and that can lead to wasted time and resources, and sometimes put patients at risk.

The integrity of a publication can be compromised in many ways. Some are unintentional: typos, transcription errors or incorrect analyses. Others are deliberate: image manipulation, data falsification and plagiarism[1]. How publication integrity was compromised is secondary to whether the paper is reliable. Unreliable data or conclusions are problems irrespective of the cause.

Enter checklists, which have helped in the structuring of complex procedures in health care and other industries. One of us (C.K.G.) helped to develop a checklist for assessing the quality of institutional investigations of researchers' conduct[2]. Academic publishing has also introduced a suite of checklists to be completed by authors when submitting papers, to make sure certain aspects of the research are fully reported.

Here we present a tool — the REAPPRAISED checklist — that aims to help readers, journal editors and anyone else assess whether a paper has flaws that call its integrity into question. We developed it on the basis of our own experience and extensive consultation with research administrators and journal editors. Although designed for clinical and animal studies, the structured approach to investigation applies more broadly. Readers, peer reviewers, journals, publishers and institutions can use it to assess whether to trust a paper's findings. Our checklist should not be confused with journals' submission checklists, which are filled in by authors before publication and indicate what items are reported in the manuscript. The REAPPRAISED checklist can be used by anyone struggling to assess a submitted or published article, and includes common-sense assessments that go beyond the text itself. It can, and should, be applied independently of whether misconduct is suspected. Its use can help to speed up the identification and correction of flawed papers, preventing wasted resources and even protecting patients from harm.

Cause for alarm

How did we come to see the need for this tool? From early 2013, three of us (A.A., A.G., M.J.B.) began to contact journals about multiple, serious problems we had identified in 33 reports of trials led by bone-health researchers Yoshihiro Sato and Jun Iwamoto[3]. The first retraction did not appear until late 2015.

This delay is all the more regrettable given that concerns had been raised more than a decade earlier. In 2003, a publication by Sato[4] assessed a very rare complication of treatment of Parkinson's disease. Within a year, a letter[5] to the editor remarked on how surprising it was that the research group had managed to identify 40 people with this complication in a very short time, because the writer's specialist institute had seen only two cases "in living memory". Others had brought up concerns about ethical oversight[6], as well as failure of randomization, implausible recruitment and outcomes, and other problems[7], but no editorial comment or correction occurred at that time; retraction

© 2020 Springer Nature Limited. All rights reserved.

THE 'REAPPRAISED' CHECKLIST FOR EVALUATION OF PUBLICATION INTEGRITY

Not all items will be applicable to every publication, and other questions might be relevant for individual categories.

R — Research governance

- ☐ Are the locations where the research took place specified, and is this information plausible?
- ☐ Is a funding source reported?
- ☐ Has the study been registered?
- ☐ Are details such as dates and study methods in the publication consistent with those in the registration documents?

E — Ethics

- ☐ Is there evidence that the work has been approved by a specific, recognized committee?
- ☐ Are there any concerns about unethical practice?

A — Authorship

- ☐ Do all authors meet criteria for authorship?
- ☐ Are contributorship statements present?
- ☐ Are contributorship statements complete?
- ☐ Is authorship of related papers consistent?
- ☐ Can co-authors attest to the reliability of the paper?

P — Productivity

- ☐ Is the volume of work reported by research group plausible, including that indicated by concurrent studies from the same group?
- ☐ Is the reported staffing adequate for the study conduct as reported?

P — Plagiarism

- ☐ Is there evidence of copied work?
- ☐ Is there evidence of text recycling (cutting and pasting text between papers), including text that is inconsistent with the study?

R — Research conduct

- ☐ Is the recruitment of participants plausible within the stated time frame for the research?
- ☐ Is the recruitment of participants plausible considering the epidemiology of the disease in the area of the study location?
- ☐ Do the numbers of animals purchased and housed align with numbers in the publication?
- ☐ Is the number of participant withdrawals compatible with the disease, age and timeline?
- ☐ Is the number of participant deaths compatible with the disease, age and timeline?
- ☐ Is the interval between study completion and manuscript submission plausible?
- ☐ Could the study plausibly be completed as described?

A — Analyses and methods

- ☐ Are the study methods plausible, at the location specified?
- ☐ Have the correct analyses been undertaken and reported?
- ☐ Is there evidence of poor methodology, including:
  - ☐ Missing data
  - ☐ Inappropriate data handling
  - ☐ '*P*-hacking': biased or selective analyses that promote fragile results
  - ☐ Other unacknowledged multiple statistical testing
- ☐ Is there outcome switching — that is, do the analysis and discussion focus on measures other than those specified in registered analysis plans?

I — Image manipulation

- ☐ Is there evidence of manipulation or duplication of images?

S — Statistics and data

- ☐ Are any data impossible?
  - ☐ Are subgroup means incompatible with those for the whole cohort?
  - ☐ Are the reported summary data compatible with the reported range?
  - ☐ Are the summary outcome data identical across study groups?
  - ☐ Are there any discrepancies between data reported in figures, tables and text?
  - ☐ Are statistical test results compatible with reported data?
- ☐ Are any data implausible?
  - ☐ Are any of the baseline data excessively similar or different between randomized groups?
  - ☐ Are any of the outcome data unexpected outliers?
  - ☐ Are the frequencies of the outcomes unusual?
  - ☐ Are any data outside the expected range for sex, age or disease?
  - ☐ Are there any discrepancies between the values for percentage and absolute change?
  - ☐ Are there any discrepancies between reported data and participant inclusion criteria?
  - ☐ Are the variances in biological variables surprisingly consistent over time?

E — Errors

- ☐ Are correct units reported?
- ☐ Are numbers of participants correct and consistent throughout the publication?
- ☐ Are calculations of proportions and percentages correct?
- ☐ Are results internally consistent?
- ☐ Are the results of statistical testing internally consistent and plausible?
- ☐ Are other data errors present?
- ☐ Are there typographical errors?

D — Data duplication and reporting

- ☐ Have the data been published elsewhere?
- ☐ Is any duplicate reporting acknowledged or explained?
- ☐ How many data are duplicate reported?
- ☐ Are duplicate-reported data consistent between publications?
- ☐ Are relevant methods consistent between publications?
- ☐ Is there evidence of duplication of figures?

© 2020 Springer Nature Limited. All rights reserved.

notices were finally issued in 2015–18.

Even after Sato admitted[8] in 2016 to making up data, only 2 of the 34 journals we contacted took the initiative to assess other papers they had published by this group.

Investigations conducted at the four institutions at which Sato and Iwamoto worked were also, in our opinion, misdirected and incomplete. They focused on identifying researcher misconduct, not publication reliability. They assessed only 90 of 351 potentially compromised publications, and only after we contacted them in 2017 with concerns[9]. Two institutions failed to reach a conclusion about the reliability of 56 publications of the 78 they considered, because they could not determine whether misconduct had occurred. The other two did not report assessments on individual papers. (The institutions maintain that their investigations were appropriate, but have not responded directly to the criticisms, and two stated that it was difficult to assess publication integrity independently of misconduct.)

So far, at least 90 papers by Sato or Iwamoto have been retracted. More than twice that number remain in the literature, including 5 of the 33 clinical-trial reports. Readers have no way of knowing whether those reports are trustworthy. Nor do they have any guidance for making their own assessments.

These papers continue to be influential. In the years since we brought our concerns to journals, those 33 clinical-trial reports have been cited more than 600 times. They have been used by other researchers as evidence to help justify at least eight clinical trials[10]. Others that were later retracted were used in an effectiveness review conducted by the US Agency for Healthcare Research and Quality. They provided the sole evidence that bisphosphonates, commonly used for osteoporosis, could prevent fractures in patients at high risk of falls[10].

Other cases had similar outcomes. More than ten years elapsed between the initial notification of concerns about publications of clinical studies in anaesthesia by Yoshitaka Fujii and the first of 183 retractions (see go.nature.com/2pvw2ax). Concerns raised about problems in a paper by Andrew Wakefield on autism and vaccination did not result in full retraction until 12 years after its publication. Flawed papers such as these influence future research initiatives, are incorporated into clinical guidelines and influence medical practice and public perceptions, so could potentially harm millions of patients.

Improving assessment

Our REAPPRAISED checklist facilitates systematic evaluation through 11 categories (see 'The 'REAPPRAISED' list for evaluation of publication integrity'). It covers ethical oversight and funding, research productivity and investigator workload, validity of randomization, plausibility of results and duplicate data reporting. It can identify problems from isolated data errors to data fabrication or falsification. Some of these questions should ideally have been asked by reviewers and editors before publication, and not all questions will apply to every paper, but it is still useful to consider all the questions collectively when assessing an article.

We developed the checklist in our roles as researchers (A.G., M.J.B., A.A.) and journal editor (A.A.K.) while evaluating thousands of publications; C.K.G., a specialist in institutional investigations, helped to refine it.

Our experience suggests that using REAPPRAISED can assist in decision-making before and after publication. A.A.K. implemented an earlier version of the checklist as editor-in-chief of *Anaesthesia*, which reviews nearly 1,000 submissions per year. Each is screened with the checklist; if concerns arise, individual patient data are requested and reviewed carefully for errors, inconsistencies or other red flags. In the past two years of routine use, editors have identified integrity problems in 42 submissions. A large subset of these — comprising work from 12 research groups — was serious enough to notify authors' institutions. Six of the 12 ensuing investigations confirmed that data had been falsified or fabricated; two laid the blame on errors; and four are ongoing (A.A.K., personal communication).

We have also found this checklist effective in communicating with journals, having used it to submit a structured list of concerns to journal editors for the 56 publications for which institutional investigations were inconclusive. That has led to 29 retractions and 5 expressions of concern.

Implementation

We would like to see the checklist used during both manuscript review and post-publication evaluation. The fact that it separates assessment of publication integrity from the investigation of research misconduct will speed up evaluations. It could even be published alongside decisions to retract, issue an expression of concern, correct a paper, or let it stand.

We expect that use of REAPPRAISED will lead to more detailed, efficient, consistent and transparent evaluations of publication integrity, and thus faster and more accurate reporting of corrections and retractions. These improvements will benefit the researchers, clinicians, policymakers, patients and others who rely on the literature to make decisions. People using the tool will be able to help refine it as they gain experience, and it will help them to develop standards to assess the integrity of publications and act accordingly. We hope others will join in our efforts to implement and refine REAPPRAISED both informally and in future publications.

Integrity problems often cluster. Authors who have a paper retracted for misconduct are more likely to accrue multiple retractions than are those whose work was retracted for other reasons[11]. Publications by their co-authors can also be compromised[12]. The checklist could help journal staff and investigation committees to decide when an assessment should be broadened to include other papers by a particular researcher and collaborators.

Ironically, a checklist that puts aside the question of misconduct might aid in evaluations of inappropriate behaviour. If multiple concerns are identified, or the concerns identified are those often associated with misconduct, the entire body of an author's work should be systematically assessed.

Publishers' integrity groups should adopt the checklist (or ask those expressing concerns to do so). Funders and government regulators should disseminate it to publishers, research institutions and other stakeholders. Peer reviewers and readers can use it on their own initiative, and those who have a nagging feeling about a publication can use it to work through their concerns and, if merited, communicate those to a journal.

If the goal is trustworthy literature, the integrity of publications — not just determination of misconduct — should be the focus of investigation.

The authors

Andrew Grey and **Mark J. Bolland** are associate professors of medicine at the University of Auckland, New Zealand. **Alison Avenell** is a professor at the Health Services Research Unit, University of Aberdeen, UK. **Andrew A. Klein** is a consultant anaesthetist at Royal Papworth Hospital, Cambridge, UK, and editor-in-chief of *Anaesthesia*. **C. K. Gunsalus** is the director of the National Center for Professional and Research Ethics at the University of Illinois at Urbana–Champaign, USA.
e-mail: a.grey@auckland.ac.nz

1. Nylenna, M. & Simonsen, S. *Lancet* **367**, 1882–1884 (2006).
2. Gunsalus, C. K., Marcus, A. R. & Oransky, I. *J. Am. Med. Assoc.* **319**, 1315–1316 (2018).
3. Bolland, M. J., Avenell, A., Gamble, G. D. & Grey, A. *Neurology* **87**, 2391–2402 (2016).
4. Sato, Y., Asoh, T., Metoki, N. & Satoh, K. *J. Neurol. Neurosurg. Psychiatry* **74**, 574–576 (2003); retraction **89**, e3 (2018).
5. Clarke, C. E. *J. Neurol. Neurosurg. Psychiatry* **75**, 510–511 (2004).
6. Caffrey, N., Casey, M. & Walsh, B. *J. Bone Miner. Res.* **21**, 491 (2006); correction **32**, 1151–1151 (2017).
7. Halbekath, J. M., Schenk, S., von Maxen, A., Meyer, G. & Mühlhauser, I. *Arch. Intern. Med.* **167**, 513–514 (2007).
8. Gross, R. A. *Neurology* **87**, 2388 (2016).
9. Grey, A., Bolland, M., Gamble, G. & Avenell, A. *Res. Integr. Peer Rev.* **4**, 3 (2019).
10. Avenell, A., Stewart, F., Grey, A., Gamble, G. & Bolland, M. J. *BMJ Open* **9**, e031909 (2019); correction **9**, e031909corr1 (2019).
11. Steen, R. G., Casadevall, A. & Fang, F. C. *PLoS ONE* **8**, e68397 (2013); correction https://doi.org/10.1371/journal.pone.0068397 (2013).
12. Carlisle, J. B. & Loadsman, J. A. *Anaesthesia* **72**, 17–27 (2017).

© 2020 Springer Nature Limited. All rights reserved.

Correspondence

Brazil's biofuel plans drive deforestation

The scientific community, Brazil's policymakers and the public need to take coordinated action against plans to amplify biofuel production at the expense of the Amazon rainforest. Brazilian President Jair Bolsonaro lifted the ban on sugar-cane cultivation in the Amazon by decree in November 2019 to help boost the country's biofuel production. Plans to produce more oil palm for biofuel in the state of Roraima follow the inauguration of a processing plant there last April. And in the west of the state of Amazonas, new roads could open up previously inaccessible areas for palm plantations and drive further devastating cycles of deforestation.

After the president took over in January 2019 (C. A. Nobre *Nature* **574**, 455; 2019), a government consortium announced investments of 4.4 billion reais (US$1.1 billion) in six Amazon states — Amazonas, Acre, Amapá, Mato Grosso, Rondônia and Roraima — for the installation of power plants fuelled by maize ethanol. Maize ethanol was chosen because of the ban on sugar-cane production, introduced to curb deforestation and the loss of ecosystem services essential to Brazil's agriculture and to mitigate global climate change (L. Ferrante and P. M. Fearnside *Science* **359**, 1476; 2018).

Lifting the ban adds to the already huge pressure on the rainforest from cattle ranching, soya-bean production, hydroelectric dams and mining.

Lucas Ferrante, **Philip M. Fearnside** National Institute of Amazonian Research (INPA), Manaus, Amazonas, Brazil.
lucasferrante@hotmail.com

Social media sows ecology consensus

All too often, social media deepens divisions rather than bridges them. But we have had a refreshingly different experience: through an open call on Twitter, we assembled a global team of more than 50 authors at different career stages and from a range of scientific disciplines. We aim to build a consensus on the drivers of ecological disturbances and their impact (see go.nature.com/2pwwrrb).

Disturbances such as wildfires, hurricanes and sea-level rise have long-lasting implications that can fundamentally alter the structure and functions of an ecosystem. Understanding their complex causes and consequences requires rigorous analysis of multi-stressor interactions. To this end, we developed an integrated concept of disturbance ecology that is independent of the scale of the investigation. This method draws together ideas from disciplines such as data science, molecular ecology and palaeoecology. Our approaches are based on input from places as ecologically varied as Argentina, Luxembourg and Singapore. They fulfil minimum reporting standards and aim to meet key requirements for understanding long-term socio-ecological impacts.

We have found that social media can stimulate open science and attract contributions from researchers, irrespective of their career stage or financial support.

Emily B. Graham Pacific Northwest National Laboratory, Richland, Washington, USA.
emily.graham@pnnl.gov

Stefan Krause University of Birmingham, UK.

In defence of the paper census

We agree that hard-copy censuses are crucial to public-health surveys in developing economies (see *Nature* **574**, 296; 2019). In our view, they should be continued and not replaced by online 'big data' collection. Furthermore, your suggestion that non-governmental organizations (NGOs) — as opposed to governments — could carry out censuses would pose several risks.

For example, NGOs would be unlikely to guarantee country-level coverage, uniformity in data collection, periodicity or commitment to conservation and oversight of data access. That could seriously impair the availability, quality, long-term observability and accessibility of complete national data sets. Collection of data online only is a limiting factor, especially for the most disadvantaged segments of the population. Although electronic gathering aided by collectors might improve coverage, it could also breach rights to privacy.

More frequent sampling, as envisaged by some countries, would require data sets to be catalogued year on year for comparative purposes, raising the possibility of missed or double-counted entries.

We urge caution. All too often, governments are willing to renege on their responsibilities to citizens — particularly if NGOs are poised to meet the challenge. The work of Roy Burstein and his colleagues in tracking child mortality (*Nature* **574**, 353–358; 2019) would not have been possible without the granularity of household-level census returns.

Ciara Breathnach, Tiziana Margaria University of Limerick, Limerick, Ireland.
ciara.breathnach@ul.ie

Consider context to avert catastrophe

Alisha Graves and her colleagues suggest that catastrophe in Africa's Sahel region could be averted by boosting food production and personal security, along with improving efforts to slow population growth (*Nature* **575**, 282–286; 2019). In my view, however, their suggestions overlook the realities of conflicts on the ground.

Increasing military spending can, as they themselves point out, aggravate the problem. Heavy-handed responses by the army, such as scorched-earth tactics and human-rights abuses, undermine communities' ability to cope with the effects of climate change and risk driving up recruitment to armed opposition groups (see https://shoring-up-stability.org/). Even national governments in the region accept the need to step away from military force and move towards governance that can tackle problems.

The authors suggest that more international aid could address the shortfall in basic services (in health and education, for example) that is resulting from increased military spending. Evidence from the region points to the contrary: failure of governance and service provision by national and local governments stoked initial grievances, which continue to drive conflict. Increased dependence on foreign aid could lock in this problem.

Bringing science and policy together can inform better responses, but these need to take the context and its people as the starting point.

Janani Vivekananda Adelphi, Berlin, Germany.
vivekananda@adelphi.de

© 2020 Springer Nature Limited. All rights reserved.

News & views

Materials science

Controllable strain enhances a semiconductor

Jian Shi

Semiconductors known as halide perovskites have remarkable optoelectronic properties, but their structural instability limits practical applications. A solution has been found that involves squeezing the compounds' crystal lattices. **See p.209**

The surprising discovery[1] in 2009 that compounds known as halide perovskites can convert sunlight into electricity triggered a revolution in photovoltaics (solar cells), inspiring new cell designs that will enable solar energy to be harnessed efficiently and at low cost. However, tuning the properties of these crystals to enable practical applications has been a long-standing challenge. On page 209, Chen *et al.*[2] report that a solution to this problem has finally been found.

Discovering new materials and identifying appropriate applications can sometimes be achieved serendipitously, but often takes decades, and has historically required centuries or even millennia. The first halide perovskite was discovered in the 1890s[3], but its potential remained untapped until a decade ago[1], when it took photovoltaics by storm. As its name suggests, halogen atoms in the compound enable the formation of a cubic (or pseudocubic) array known as a perovskite structure.

Harnessing the full potential of halide perovskites for technological applications has been difficult. A major obstacle is the tendency of one of the best-performing perovskite crystals, α-formamidinium lead iodide ($HC(NH_2)_2PbI_3$, known as α-$FAPbI_3$), to assume a hexagonal structure at room temperature (Fig. 1a) – the approximate temperature at which photovoltaic devices operate. This hexagonal structure cannot respond to most of the frequencies of light in solar radiation, and hence is not of interest for technological applications. It would therefore be helpful to stabilize the structure of α-$FAPbI_3$.

Several strategies can be used to engineer a material's properties. Two of the most effective are to alter the material's composition or to use it at high or low temperatures, but the costs involved for commercial applications can be tremendous. Scientists have therefore developed a simple but extremely useful approach known as strain engineering, which has been used to tune the electronic properties of semiconductors.

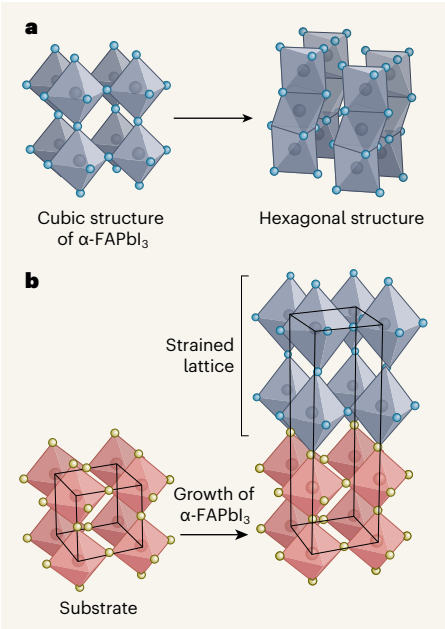


Figure 1 | The stabilization and strain engineering of a semiconductor. a, The semiconductor α-formamidinium lead iodide ($HC(NH_2)_2PbI_3$, known as α-$FAPbI_3$) belongs to the halide perovskite family of compounds, and has potential applications in solar cells. However, its cubic perovskite lattice is unstable, and converts to a hexagonal form that is unsuitable for practical applications. The octahedra represent subunits of the lattices: iodine atoms at the vertices surround a central lead atom; $(HC(NH_2)_2)^+$ ions fill the gaps between octahedra, but are not shown. **b**, Chen *et al.*[2] report that the structure of α-$FAPbI_3$ can be stabilized by growing it on a stable halide perovskite (the substrate) that has an analogous structure, so that the atoms in the two lattices align. Because the lattice dimensions of the substrate are smaller than those of α-$FAPbI_3$, the crystal lattice of the latter is squeezed (put under strain). This strain increases the mobility of electrical charge carriers in the material.

When a crystal is compressed or stretched, the resulting deformation is called strain. Strain is calculated by dividing the change of length of a deformed object by the object's original length, and is expressed as a percentage of the original length. In semiconductors, strain can alter the mobility of charge carriers – a property that characterizes how fast a charge carrier, such as an electron, can travel in a crystal subjected to an electric field. By changing the charge mobility, the electronic properties of a semiconductor can be altered.

Semiconductors in modern electronic devices are most commonly used as thin films, and the maximum sustainable strain of a film is less than a few per cent[4]. Nevertheless, strain modulation can be a highly effective tool for enhancing electronic properties. One such success story is that of lasers based on layered semiconductor structures known as quantum wells. By straining the quantum-well layers, the electrical current needed to power a laser can be reduced by as much as ten times, thus improving the energy efficiency of the devices[5,6]. Most commercial quantum-well lasers are therefore strained[5].

One of the most interesting but problematic features of halide perovskites is that only a few, including α-$FAPbI_3$, have high charge mobility and absorb light strongly, but the technologically useful crystal structures of these few compounds are unstable – some can spontaneously transform into other phases in less than a second. In photovoltaic devices, fast-moving charge carriers and strong light absorption are both needed to convert solar energy into electricity with high efficiency. Unfortunately, structural stability, high charge mobility and the ability to absorb light strongly don't seem to coexist in perovskite halides.

Chen *et al.* have therefore used strain engineering to tackle this problem. They grew crystalline α-$FAPbI_3$ from a solution so that it formed on another, more stable halide perovskite (the substrate). The $FAPbI_3$ atoms in the growing crystal align with the cubic structure of the atoms in the substrate, thereby forming a pseudocubic structure themselves (Fig. 1b). The alignment of atoms in different materials is called epitaxy. This epitaxy locks α-$FAPbI_3$ into the pseudocubic structure as a result of the strong chemical forces between it and the substrate, preventing its transformation into the

© 2020 Springer Nature Limited. All rights reserved.

undesirable hexagonal structure. The authors find that the pseudocubic structure remains stable for at least a year at room temperature.

A compressive strain is imposed on the α-$FAPbI_3$ film because the dimensions of the cubic array of the substrate are different from those of the natural atomic array of α-$FAPbI_3$. Chen and colleagues were therefore able to control the strain of α-$FAPbI_3$ from 0 to 2.4% compressive deformation by growing $FAPbI_3$ on substrates that have different lattice dimensions. The authors found that this squeezing of the α-$FAPbI_3$ crystal increases the mobility of positively charged quasiparticles called holes, which correspond to the absence of electrons in the crystal. The authors attribute this increased mobility to the modification of the electronic structure of the crystal under compressive strain: compression leads to faster oscillations of the holes' wavefunctions, speeding up the movement of charge wavepackets (superpositions of wavefunctions) and thus producing higher charge mobility.

Previous work on the strain engineering of halide perovskite films lacked strain control[7] or involved straining methods that are harder to use[8,9]. By contrast, Chen and colleagues' study provides an extremely accessible and practical avenue through which to explore and use the physical properties of strained halide perovskites.

Questions remain about how the authors' findings could be used in solar cells. Currently available halide perovskite photovoltaic devices do not contain a genuinely epitaxial substrate, and so new cell designs will be needed to make use of the reported discovery. But a range of halide perovskite compounds are available that have similar atomic arrays to α-$FAPbI_3$, and which exhibit many different technologically important electronic properties. Chen and co-workers' study therefore suggests that there is plenty of scope for designing and developing epitaxial quantum-well devices using these materials, by mimicking the way in which quantum-well devices were developed using semiconductors from the III–V family of materials. This might bring down the cost of manufacturing these devices.

Finally, it will be interesting to see whether crystals of halide perovskites can be grown with sufficient atomic precision to make superlattices — periodic structures that contain multiple layers of two or more materials. The use of halide perovskites in superlattices could open up otherwise inaccessible electronic band structures, thereby allowing a rich array of physics to be explored, and emerging quantum-well devices to be further developed.

Jian Shi is in the Department of Materials Science and Engineering, Rensselaer Polytechnic Institute, Troy, New York 12180, USA.
e-mail: shij4@rpi.edu

1. Kojima, A. *et al. J. Am. Chem. Soc.* **131**, 6050–6051 (2009).
2. Chen, Y. *et al. Nature* **577**, 209–215 (2020).
3. Wells, H. L. *Z. Anorg. Allg. Chem.* **3**, 195–210 (1893).
4. Li, J. *et al. MRS Bull.* **39**, 108–114 (2014).
5. Adams, A. R. *IEEE J. Sel. Top. Quantum Electron.* **17**, 1364–1373 (2011).
6. Yablonovitch, E. & Kane, E. O. *J. Lightw. Technol.* **4**, 504–506 (1986).
7. Zhu, C. *et al. Nature Commun.* **10**, 815 (2019).
8. Wang, Y. *et al. Phys. Rev. Mater.* **2**, 076002 (2018).
9. Wang, Y. *et al. Sci. Adv.* **4**, eaar3679 (2018).

Population biology

Predator–prey cycles achieved at last

Alan Hastings

A combination of laboratory experiments and mathematical and statistical analysis provides an affirmative answer to a decades-old question — can a predator and its prey coexist indefinitely? **See p.226**

A key question in ecology is what allows species to persist over time — particularly when there are pairs of species in which one is an exploiter and the other its victim. A long-standing theory attempts to answer this question by explaining how relative numbers of predators and their prey can cycle continuously[1]. First, prey numbers would increase, giving the predator more food. The subsequent increase in predators would lead to a decline in prey. Predator numbers would then decline owing to a lack of food, restarting the cycle. However, it has proved unexpectedly challenging to demonstrate this type of persistent predator–prey cycle in simple controlled systems in the laboratory. On page 226, Blasius *et al.*[2] report just such a demonstration, succeeding where almost 90 years of experimental work has failed.

Ecological theories of persistent predator–prey cycles are supported by the apparent existence of such cycles in nature, for instance between the lynx and hare in Canada[3]. However, it is hard to prove that these cycles endure in the wild, because observations over many decades would be needed. But if the theories provide a complete explanation of natural cycles, then it should be possible to demonstrate persistent cyclic behaviour in the laboratory, using species that have much shorter cycle times.

The challenge posed by such a demonstration was exemplified in 1934 by the ecologist Georgii Gause[4], who studied the dynamics of two unicellular organisms — the predator *Didinium nasutum* and its prey, *Paramecium caudatum*. Gause found that, on the one hand, if the predator was efficient, it ate up all the prey and then starved. On the other hand, if part of the environment helped to conceal the prey, the predator was less efficient — and so starved (Fig. 1a). Coexistence and long-term cycles could be achieved only through artificial means — namely, by adding prey at regular intervals.

In 1974, work with the same system showed that, by making the predator less efficient and by providing the prey with less food, the two populations could persist for longer[5]. Even so, coexistence could be maintained for just a few predator–prey cycles. Since then, some models that allow long-term cycle persistence have focused on space, for instance incorporating metapopulation dynamics[6]. In this phenomenon, subpopulations of a species migrate around a larger region. Although a subpopulation might become extinct in one area, the species persists across the region as a whole and can migrate back into that area. However, a better understanding of whether exploiter–victim cycles can persist locally without external input is still sorely needed.

Blasius and colleagues studied the aquatic invertebrate *Brachionus calyciflorus* and its prey, the green algal species *Monoraphidium minutum*. They found that, under simple and constant environmental conditions, the two species could coexist for more than a year — that is, over 50 predator–prey cycles. This result finally demonstrates that the long-standing theory of persistent cycles can be consistent with the reality of simple ecological systems.

Next, the researchers carried out a rigorous statistical analysis of the cycle dynamics in their system. Specifically, they used wavelet analysis, which focuses on dynamics over short periods; the technique has become a standard way to study the presence of periodic behaviour in ecological time series[7]. The analysis revealed interesting dynamic phenomena. The oscillations in the relative numbers of each species showed a characteristic lag in phase, with predator numbers mostly changing in the wake of altered prey numbers.

© 2020 Springer Nature Limited. All rights reserved.

https://doi.org/10.1038/d41586-019-03603-3

News & views

Population biology

Predator–prey cycles achieved at last

Alan Hastings

A combination of laboratory experiments and mathematical and statistical analysis provides an affirmative answer to a decades-old question — can a predator and its prey coexist indefinitely?

A key question in ecology is what allows species to persist over time — particularly when there are pairs of species in which one is an exploiter and the other its victim. A long-standing theory attempts to answer this question by explaining how relative numbers of predators and their prey can cycle continuously[1]. First, prey numbers would increase, giving the predator more food. The subsequent increase in predators would lead to a decline in prey. Predator numbers would then decline owing to a lack of food, restarting the cycle. However, it has proved unexpectedly challenging to demonstrate this type of persistent predator–prey cycle in simple controlled systems in the laboratory. Writing in *Nature*, Blasius *et al.*[2] report just such a demonstration, succeeding where almost 90 years of experimental work has failed.

Ecological theories of persistent predator–prey cycles are supported by the apparent existence of such cycles in nature, for instance between the lynx and hare in Canada[3]. However, it is hard to prove that these cycles endure in the wild, because observations over many decades would be needed. But if the theories provide a complete explanation of natural cycles, then it should be possible to demonstrate persistent cyclic behaviour in the laboratory, using species that have much shorter cycle times.

The challenge posed by such a demonstration was exemplified in 1934 by the ecologist Georgii Gause[4], who studied the dynamics of two unicellular organisms — the predator *Didinium nasutum* and its prey, *Paramecium caudatum*. Gause found that, on the one hand, if the predator was efficient, it ate up all the prey and then starved. On the other hand, if part of the environment helped to conceal the prey, the predator was less efficient — and so starved (Fig. 1a). Coexistence and long-term cycles could be achieved only through artificial means — namely, by adding prey at regular intervals.

In 1974, work with the same system showed that, by making the predator less efficient and by providing the prey with less food, the two populations could persist for longer[5]. Even so, coexistence could be maintained for just a few predator–prey cycles. Since then, some models that allow long-term cycle persistence have focused on space, for instance incorporating metapopulation dynamics[6]. In this phenomenon, subpopulations of a species migrate around a larger region. Although a subpopulation might become extinct in one area, the species persists across the region as a whole and can migrate back into that area. However, a better understanding of whether exploiter–victim cycles can persist locally without external input is still sorely needed.

Blasius and colleagues studied the aquatic invertebrate *Brachionus calyciflorus* and its prey, the green algal species *Monoraphidium minutum*. They found that, under simple and constant environmental conditions, the two species could coexist for more than a year — that is, over 50 predator–prey cycles. This result finally demonstrates that the long-standing theory of persistent cycles can be consistent with the reality of simple ecological systems.

Next, the researchers carried out a rigorous statistical analysis of the cycle dynamics in their system. Specifically, they used wavelet analysis, which focuses on dynamics over short periods; the technique has become a standard way to study the presence of periodic behaviour in ecological time series[7]. The analysis revealed interesting dynamic phenomena. The oscillations in the relative numbers of each species showed a characteristic lag in phase, with predator numbers mostly changing in the wake of altered prey numbers. However, these oscillations would undergo sudden shifts, without any external driver. During these transient periods, both populations would oscillate out of sync with one another, before the in-phase cyclic dynamic resumed (Fig. 1b).

The authors also performed experiments in which they introduced pulses of nutrients to the species' environment. This mimics seasonal environmental changes experienced in many natural systems. The overall dynamics of the predator–prey cycle remained the same in these conditions, indicating that predator–prey interactions can govern cycle dynamics, even in a seasonally varying environment.

Although Blasius and colleagues' work

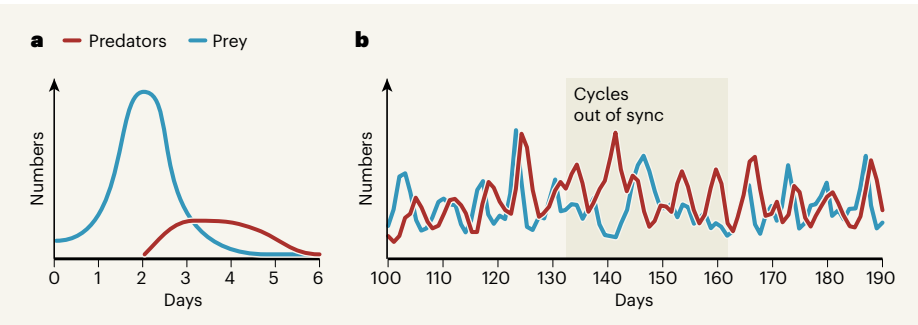


Figure 1 | A difficult balance to strike. Ecological theory predicts that the relative numbers of a predator and its prey should oscillate in persistent cycles, but demonstrating this in the laboratory has been hard. **a**, Classic experiments[4] from the 1930s failed to achieve cycles, because predators were either too inefficient at eating prey (not shown) or too efficient, and ate them all. Either way, predator numbers dwindled. **b**, Blasius *et al.*[2] have succeeded in achieving persistent predator–prey cycles in the laboratory. They found that cycles were mostly in phase, with predator cycles slightly lagging behind prey cycles. However, there were transient periods in which the two — for unknown reasons — fell out of sync. (Adapted from Fig. 2b of the paper.)

© 2019 Springer Nature Limited. All rights reserved.

answers one question, it raises others. First and foremost, why are the results so different from those of earlier experiments? It is not just a quirk of the species used, because the authors repeated the experiment using a different prey species and achieved the same result. One contributing factor could be that the researchers used a predator that has a complex life history involving transitions through several stages, from egg to adulthood. Another could be that persistence depends on the specific amplitude of oscillation in the cycle.

In support of these ideas, previous work involving a predator with a simpler life cycle demonstrated that populations of predators, prey or both went extinct if cycles reached an amplitude at which population numbers become low[4]. In addition, analysis of a different marine predator–prey combination[8] showed that cyclic dynamics eventually gave way either to extinction or to persistence of the two populations in equilibrium. And mathematical theory[1] suggests that the small-amplitude cycles observed by Blasius and colleagues would occur only under very specific conditions. Demonstrating persistent cycles in other systems is therefore a key challenge for the future. The life-history characteristics of the predator used by Blasius and colleagues could be a starting point in the search for systems that show similar behaviour.

Another intriguing avenue for future study is to determine the relationship between transient dynamics and switches in system behaviour. Does a sudden change in the dynamics of a cycling system reflect changing environmental conditions, or is it an intrinsic phenomenon? The existence of unexpected, transient shifts in predator–prey dynamics limits the predictability of ecological systems[9]. In the real world, sudden shifts of this nature could spur an erroneous quest to identify drivers of the change, whereas the current study indicates that there may be none. The possibility of transient dynamics could also make it more difficult to predict the effect of ecological-management strategies.

Blasius *et al.* have provided one of the clearest demonstrations so far of transient dynamics in the laboratory, where it is absolutely certain that no external influences caused the change. It will be important to bring together this work (and another convincing experimental demonstration[10]) with theoretical models that are currently being developed[9]. Doing so should provide more insights into the role of transient dynamics in natural systems.

Alan Hastings is in the Department of Environmental Science and Policy, University of California, Davis, Davis, California 95616, USA, and at the Santa Fe Institute, Santa Fe, New Mexico, USA.
e-mail: amhastings@ucdavis.edu

1. Kendall, B. E. *et al. Ecology* **80**, 1789–1805 (1999).
2. Blasius, B., Rudolf, L., Weithoff, G., Gaedke, U. & Fussmann, G. F. *Nature* https://doi.org/10.1038/s41586-019-1857-0 (2019).
3. Elton, C. & Nicholson, M. *J. Anim. Ecol.* **11**, 215–244 (1942).
4. Gause, G. F. *The Struggle for Existence* (Williams & Wilkins, 1934).
5. Luckinbill, L. *Ecology* **55**, 1142–1147 (1974).
6. Hanski, I. *Metapopulation Ecology* (Oxford Univ. Press, 1999).
7. Cazelles, B. *et al. Oecologia* **156**, 287–304 (2008).
8. Fussmann, G. F., Ellner, S. P., Shertzer, K. W. & Hairston, N. G. *Science* **290**, 1358–1360 (2000).
9. Hastings, A. *et al. Science* **361**, eaat6412 (2018).
10. Cushing, J. & Desharnais, C. *J. Anim. Ecol.* **67**, 298–306 (1998).

© 2019 Springer Nature Limited. All rights reserved.

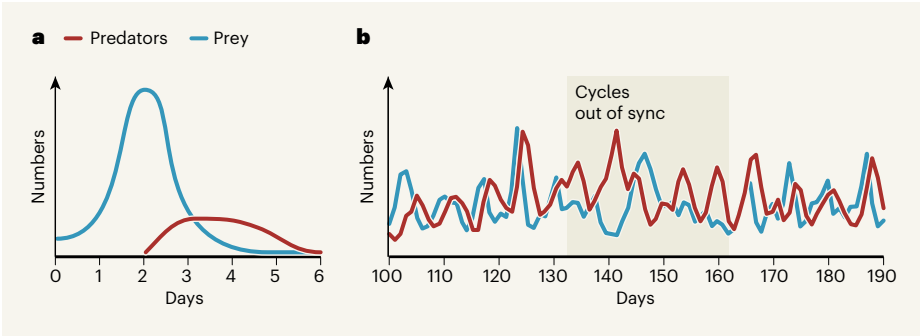


Figure 1 | A difficult balance to strike. Ecological theory predicts that the relative numbers of a predator and its prey should oscillate in persistent cycles, but demonstrating this in the laboratory has been hard. **a**, Classic experiments[4] from the 1930s failed to achieve cycles, because predators were either too inefficient at eating prey (not shown) or too efficient, and ate them all. Either way, predator numbers dwindled. **b**, Blasius *et al.*[2] have succeeded in achieving persistent predator–prey cycles in the laboratory. They found that cycles were mostly in phase, with predator cycles slightly lagging behind prey cycles. However, there were transient periods in which the two — for unknown reasons — fell out of sync. (Adapted from Fig. 2b of the paper.)

However, these oscillations would undergo sudden shifts, without any external driver. During these transient periods, both populations would oscillate out of sync with one another, before the in-phase cyclic dynamic resumed (Fig. 1b).

The authors also performed experiments in which they introduced pulses of nutrients to the species' environment. This mimics seasonal environmental changes experienced in many natural systems. The overall dynamics of the predator–prey cycle remained the same in these conditions, indicating that predator–prey interactions can govern cycle dynamics, even in a seasonally varying environment.

Although Blasius and colleagues' work answers one question, it raises others. First and foremost, why are the results so different from those of earlier experiments? It is not just a quirk of the species used, because the authors repeated the experiment using a different prey species and achieved the same result. One contributing factor could be that the researchers used a predator that has a complex life history involving transitions through several stages, from egg to adulthood. Another could be that persistence depends on the specific amplitude of oscillation in the cycle.

In support of these ideas, previous work involving a predator with a simpler life cycle demonstrated that populations of predators, prey or both went extinct if cycles reached an amplitude at which population numbers become low[4]. In addition, analysis of a different marine predator–prey combination[8] showed that cyclic dynamics eventually gave way either to extinction or to persistence of the two populations in equilibrium. And mathematical theory[1] suggests that the small-amplitude cycles observed by Blasius and colleagues would occur only under very specific conditions. Demonstrating persistent cycles in other systems is therefore a key challenge for the future. The life-history characteristics of the predator used by Blasius and colleagues could be a starting point in the search for systems that show similar behaviour.

Another intriguing avenue for future study is to determine the relationship between transient dynamics and switches in system behaviour. Does a sudden change in the dynamics of a cycling system reflect changing environmental conditions, or is it an intrinsic phenomenon? The existence of unexpected, transient shifts in predator–prey dynamics limits the predictability of ecological systems[9]. In the real world, sudden shifts of this nature could spur an erroneous quest to identify drivers of the change, whereas the current study indicates that there may be none. The possibility of transient dynamics could also make it more difficult to predict the effect of ecological-management strategies.

Blasius *et al.* have provided one of the clearest demonstrations so far of transient dynamics in the laboratory, where it is absolutely certain that no external influences caused the change. It will be important to bring together this work (and another convincing experimental demonstration[10]) with theoretical models that are currently being developed[9]. Doing so should provide more insights into the role of transient dynamics in natural systems.

Alan Hastings is in the Department of Environmental Science and Policy, University of California, Davis, Davis, California 95616, USA, and at the Santa Fe Institute, Santa Fe, New Mexico, USA.
e-mail: amhastings@ucdavis.edu

1. Kendall, B. E. *et al. Ecology* **80**, 1789–1805 (1999).
2. Blasius, B., Rudolf, L., Weithoff, G., Gaedke, U. & Fussmann, G. F. *Nature* **577**, 226–230 (2020).
3. Elton, C. & Nicholson, M. *J. Anim. Ecol.* **11**, 215–244 (1942).
4. Gause, G. F. *The Struggle for Existence* (Williams & Wilkins, 1934).
5. Luckinbill, L. *Ecology* **55**, 1142–1147 (1974).
6. Hanski, I. *Metapopulation Ecology* (Oxford Univ. Press, 1999).
7. Cazelles, B. *et al. Oecologia* **156**, 287–304 (2008).
8. Fussmann, G. F., Ellner, S. P., Shertzer, K. W. & Hairston, N. G. *Science* **290**, 1358–1360 (2000).
9. Hastings, A. *et al. Science* **361**, eaat6412 (2018).
10. Cushing, J. & Desharnais, C. *J. Anim. Ecol.* **67**, 298–306 (1998).

This article was published online on 18 December 2019.

Materials science

The physics of ice skating

Daniel Bonn

The slipperiness of ice is poorly understood at a microscopic level. Experiments that probe how the surface of ice melts and flows in response to wear help to explain the exceptionally low friction that underpins winter sports.

It is widely thought that ice skating is enabled by the formation of a layer of water on the surface of ice, which lubricates the skate. Writing in *Physical Review X*, Canale *et al.*[1] report that, although there is indeed a layer of something, it is not simply water — the lubricating layer has surprising flow properties that are very different from those of bulk water, and are intermediate between those of liquid water and ice.

The idea that a thin film of meltwater wets the surface of ice has been accepted since the nineteenth century[2]. This layer should be present beneath skates or other objects sliding on ice. Most of the debate about this topic, therefore, deals with the origin of the layer — is it the result of surface melting, heating associated with friction generated by the skate, or pressure-induced melting? Pressure melting has largely been discounted, because this process is impossible for ice below about −20 °C, whereas skating is possible at such low temperatures. Surface melting and frictional heating also seem unlikely explanations because these phenomena are not specific to solid water ice, yet ice is almost the only

© 2020 Springer Nature Limited. All rights reserved.

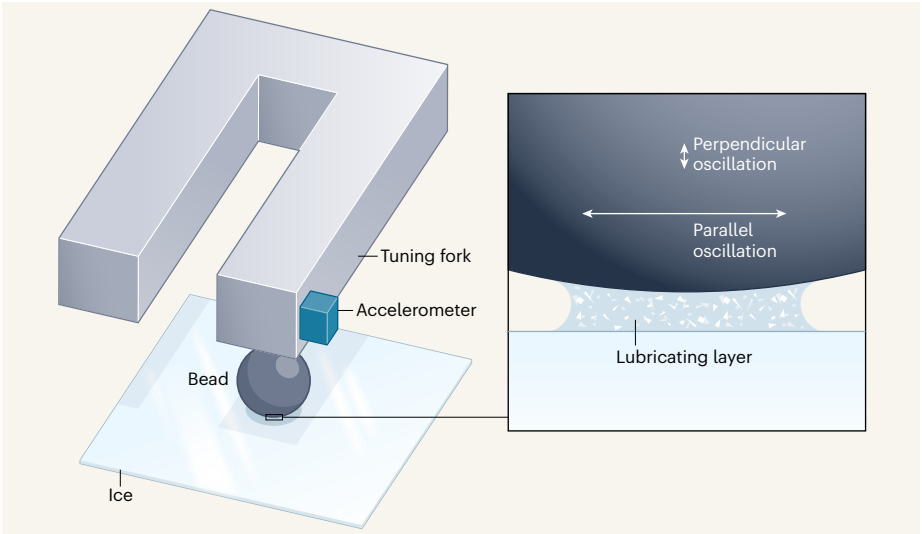


Figure 1 | Probing the lubricating layer of ice. Canale *et al.*[1] investigated the thin layer of melt water that lubricates the slipping of objects on ice. The authors attached a millimetre-scale glass bead to a prong of a macroscopic aluminium tuning fork. The fork was made to vibrate, causing the bead to oscillate over the ice surface. An accelerometer on the fork measured the amplitude of the bead's oscillations parallel and perpendicular to the surface. From these measurements, the friction between the bead and the ice, and the flow properties (rheology) of the lubrication layer, could be calculated. The observed viscosity and elasticity of the layer suggest that it consists of a mixture of water and ice particles.

material that it is possible to skate on.

Also unknown are the mechanical properties of the meltwater layer. Whether a liquid lubricates or not is mostly determined by its resistance to being squeezed out from the gap between the two rubbing surfaces — thick grease is usually needed for good lubrication. Water is not a good lubricant, because its low viscosity means that it is easily squeezed out of gaps. The idea that a layer of water is sufficient to lubricate a skate on ice is therefore strange. It doesn't even make intuitive sense, given that it is impossible to skate on a road or a kitchen floor with a layer of water on it.

Canale *et al.* report clever experiments that investigate the microscopic mechanisms responsible for the slipperiness of ice (Fig. 1). The authors measured the coefficient of friction (a measure of the friction produced at a surface) of ice and the properties of the lubrication layer simultaneously, using an experimental set-up that resembles a large tuning fork. The fork was made to vibrate, so that a millimetre-scale glass bead attached to one of its prongs oscillated across an ice surface. The bead thus functioned as a tiny ice skate, gliding for distances of the order of tens of micrometres across the same region of ice.

An accelerometer attached to the same prong of the fork as the bead measured the amplitude of the bead's oscillations parallel to the surface, and compared them with the amplitude of the driving force. Canale *et al.* used the difference in amplitude to calculate the friction force between the bead and the ice. Simultaneously, tiny oscillations of the bead (restricted to just a few tens of nanometres, so as not to influence the friction measurements) perpendicular to the ice surface allowed the authors to probe the properties of the lubrication layer. The authors used a similar procedure to the one used to measure parallel movements to measure the forces acting on the bead perpendicular to the surface. The coefficient of friction between glass and ice could then be calculated from the ratio of the two measured forces.

Canale and colleagues' procedure not only measured the amplitude change of the perpendicular movements, but also the time taken for the driving force to produce these oscillations of the bead (the phase shift). This allowed the authors to measure the mechanical impedance of the lubricating layer on the ice — a measure of its resistance to forces applied by the bead. These measurements, in turn, correlate with the flow properties (rheology) of the lubrication layer.

Taken together, the authors' measurements allowed the friction and rheology of the lubrication layer to be probed simultaneously. The experiments show that the lubrication layer has both viscous and elastic responses to an object sliding on top of it. This is not only different from the behaviour of bulk water, but also to that of ice. The results, therefore, reveal that skating requires more than just a layer of water.

So what is happening? Sophisticated spectroscopy techniques have previously been used to study the surface of ice at different temperatures[3], and revealed that melting occurs discontinuously through successive bilayers of ice molecules — the first bilayer melts at about −70 °C, and the second one at −20 °C. Friction measurements over that same temperature range have shown, however, that friction at the surface doesn't change as the bilayers melt[4]. It therefore seems that the slipperiness of ice does not just depend on its surface properties during melting, but also on its interaction with the skate.

Canale and co-workers' experiments support that idea. The authors suggest that repeated sliding over the same spot generates a mixture of ice and water, which displays both elastic behaviour (from the ice) and viscous behaviour (from the water) in response to a load. The resulting layer of material would be more difficult to squeeze out of gaps than ordinary water. This could — at least in part — explain the layer's excellent lubrication properties.

The new study also provides the beginning of an explanation of why skating is so specific to ice. Lubricated friction and concomitant wear often lead to the formation of a layer of different material, known as a third body, between the two rubbing objects. Because this process relies on the specific ways in which the surfaces of rubbing objects wear away, and often even on friction-induced chemistry, the formation of a third body is largely system-specific. In other words, few materials can form a viscoelastic, liquid–solid third body in response to friction and wear, such as that observed in Canale and colleagues' experiments.

Interesting questions remain. For example, does the observed layer form rapidly enough on pristine ice to lubricate on contact? And how does the presence of the layer explain that the optimal ice temperature for skating[4,5] is about −7 °C? (The temperature at which all speed-skating rinks are kept.) Beyond ice skating, detailed measurements of the properties of thin lubrication films are likely to be useful, because many such films do not show the same properties as the bulk lubricant[6], and because the origin and effects of the third body in lubrication remain a largely open question. These are important issues because lubrication is often the only way to reduce friction, which is responsible for an estimated 20% of the world's energy consumption[7].

Daniel Bonn is at the Institute of Physics, University of Amsterdam, Science Park 904, NL-1098XH Amsterdam, the Netherlands.
e-mail: d.bonn@uva.nl

1. Canale, L. *et al. Phys. Rev. X* **9**, 041025 (2019).
2. Faraday, M. *XXIV. Phil. Mag. Ser. 4* **17**, 162–169 (1859).
3. Sánchez. M. A. *et al. Proc. Natl Acad. Sci. USA* **114**, 227–232 (2017).
4. Weber, B. *et al. J. Phys. Chem. Lett.* **9**, 2838–2842 (2019).
5. de Koning, J. J. *et al. J. Biomech.* **25**, 565–571 (1992).
6. Dowson, D. & Higginson, G. R. *Elasto-hydrodynamic Lubrication* 1–14 (Pergamon, 1966).
7. Holmberg, K. & Erdemir, A. *FME Trans.* **43**, 181–185 (2015).

This article was published online on 16 December 2019.

© 2020 Springer Nature Limited. All rights reserved.

Neuroscience

Internal brain states in motion

Ethan Scott

The brain fluctuates between different internal states, each of which drives particular behaviours. Brain-wide imaging reveals the internal states that help zebrafish larvae to choose between exploring and hunting. **See p.239**

Internal brain states modulate how we perceive, feel about and respond to the stimuli around us. The neurons and networks that encode these various states remain mysterious, owing to the brain's enormous complexity. As a result, researchers have only a vague sense of how animals, including humans, use recent experience and context to select appropriate behaviours. On page 239, Marques *et al.*[1] perform a first-of-its-kind exploration of brain-wide activity in zebrafish larvae as the animals encounter a classic state-driven behavioural choice: whether to explore their environment or to exploit resources in the local area. By monitoring individual neurons across the brain as these decisions play out, the authors open a window onto the neural underpinnings of internal states and how they can influence behaviour.

Advanced microscopy is revolutionizing the ways in which we observe the functioning brain. Historically, neural activity has been monitored either using brain-wide methods that cannot resolve individual neurons, or with electrophysiology, which records individual neurons or small localized populations. Neither approach can address activity across vast assemblies of individual neurons – which is exactly what our brains are.

Genetically encoded calcium indicators (GECIs) offer a possible solution to this problem. These proteins provide a pulse of fluorescence when a neuron is active. When expressed throughout the brain and combined with the right microscope, they produce a blinking light field that can reveal the activity of thousands of individual neurons simultaneously.

Zebrafish larvae are small and transparent, making it possible to image GECIs across their entire brains at cellular resolution[2,3]. However, microscopy techniques have typically required that the animals are stationary, either paralysed or with their heads embedded in agarose, hampering our ability to analyse how sensory processing relates to natural behaviour.

Brain-wide GECI imaging during free-swimming behaviour became possible when the group behind the current study introduced a tracking fluorescence microscope[4].

"The authors have delivered a satisfying scheme for how exploration and exploitation behaviours are controlled by internal brain states."

This machine constantly monitors and predicts the movements of a free-swimming larva, and moves the imaging platform to compensate for the animal's movement, keeping the brain in the microscope's field of view. Meanwhile, it uses brief pulses of light to record fluorescence signals from GECIs in the brain. After spatial alignment of the images across the experiment, the approach produces brain-wide, cell-level data on neural activity, complemented by behavioural information on the animal's position, speed and movements.

In the current study, Marques and colleagues used this platform to explore the neural underpinnings of two larval behaviours that involve free swimming: exploration (efficient swimming over distance) and exploitation (localized predation). Exploration and exploitation place opposing pulls on diverse animals[5,6] – the former offers the possibility of resource discovery, but at a metabolic cost, and the latter is an effective approach for predatory feeding, but only if prey are present. These behavioural modes are influenced by external inputs and hunger level, but also alternate spontaneously, suggesting that internal brain states have a role in which behaviours are chosen.

The authors recorded videos of free-swimming larvae, and categorized each of the animals' movements to generate thorough descriptions of behaviours during exploration and exploitation phases. They showed that movements such as forward swimming and routine turns tend to occur at a similar time and indicate that the larva is in an exploratory phase. Other movements, including targeting turns and predatory strikes typical of hunting, identify exploitative phases.

Marques *et al.* next performed GECI

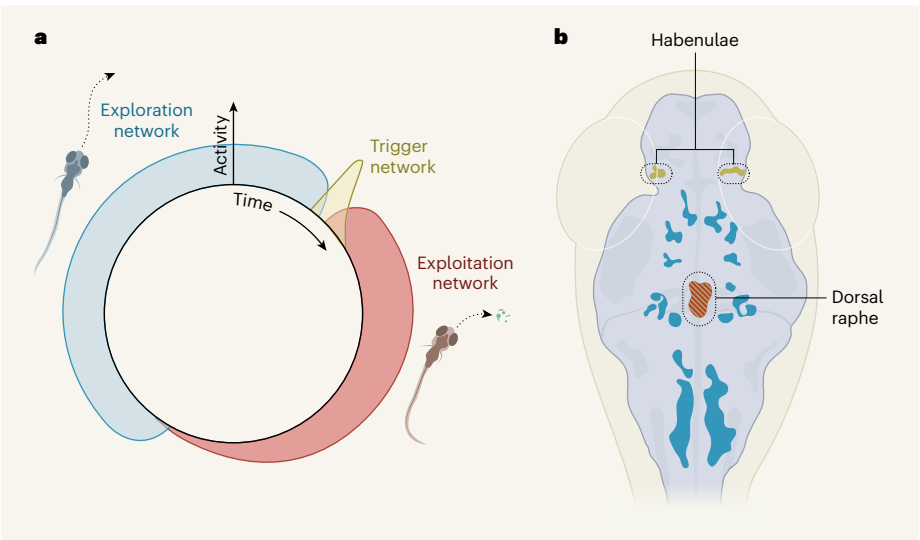


Figure 1 | Mapping exploration and exploitation. a, Using whole-brain imaging of neuronal activity in free-swimming zebrafish larvae, Marques *et al.*[1] have identified distinct categories of neuron that are active during phases of larval behaviour involving exploration (blue larva) and exploitation (hunting; red). As this cyclical graph shows, a network of neurons (blue) is steadily active throughout exploration. This activity is replaced by a peak in a 'trigger network' of neurons (yellow), which then drives the activity of exploitation-network neurons (red). Exploitation-network activity tails off over time, reaching a baseline as the exploration network becomes active once more and the cycle begins again. **b**, The exploration network is distributed throughout pre-motor regions of the zebrafish hindbrain. The exploitation network is found mainly in the dorsal raphe. The trigger network is also found in this region, and in the habenulae.

© 2020 Springer Nature Limited. All rights reserved.

imaging to analyse which brain networks encode the corresponding internal states, and what changes occur during the transitions between phases of behaviour. They extracted activity data from tens of thousands of individual neurons per animal, seeking neurons with activity that correlated with specific swimming manoeuvres, with exploration or exploitation phases, or with transitions between exploration and exploitation. This analysis revealed three notable categories of neuron (Fig. 1).

First, there are neurons dispersed across the brain that are active during exploration. Because they are broadly distributed and their activity does not show obvious peaks, they are probably involved in executing exploratory manoeuvres, rather than encoding or triggering the state itself.

Second, there is a neuronal population that is present in high numbers in a structure called the dorsal raphe. These neurons are powerfully active at the beginning of an exploitation phase. Their activity gradually weakens over time, settling to a baseline during the phase transition back to exploration.

Third, there is a 'trigger network' of neurons, located in multiple localized regions across the brain. The activity of these neurons peaks sharply when the larva is transitioning from exploration to exploitation. The strength of the peak from the trigger network predicts the strength of the subsequent activation in the dorsal raphe's exploitation-network neurons. This, in turn, predicts the duration of the ensuing exploitation phase.

How does this arrangement help to deliver appropriate behaviours at appropriate times? The expectation is that exploration and exploitation phases are influenced both by hunger and by the local availability of prey. A hungry larva should stay where it is and exploit prey when they are present, and should look for prey if they are absent locally. In agreement with this theory, the authors found that exposing larvae to light after a period of dark — revealing the presence of local prey to the larvae, whose predation is largely visual — caused increased trigger-network activity, activation of exploitation neurons in the dorsal raphe, and predation. By contrast, however, they found only a weak trend towards exploitation in hungry animals presented with prey. These mixed results provide a strong incentive to study the combinations of environmental stimuli and internal cues that spur trigger-network activity, and therefore feeding.

In summary, Marques *et al.* have delivered a plausible and satisfying scheme for how exploration and exploitation behaviours are controlled by internal brain states. Rather than being a conclusion, this is a departure point for further exploration of the underlying networks. Although whole-brain GECI imaging is a powerful tool, it provides little information on neuronal connectivity, and reveals only correlations between neurons' activity rather than the causal relationships among them.

Going forward, ablations or light-based optogenetic manipulations of targeted neurons could help to reveal exactly how activity patterns or network-level properties drive transitions between exploration and exploitation states. Furthermore, mapping out neurons' functional properties against their physical architecture (as described in a recent atlas of cell morphologies in zebrafish[7]) or connectivity (as inferred from brain-wide electron-microscopy data[8]) would provide an anatomical framework for the flow of information through the system. Such functional and anatomical information will allow researchers both to test Marques and colleagues' model of internal states, and to identify the fine details of how exploration, exploitation and trigger networks interact to control behaviour.

Ethan Scott is at the Queensland Brain Institute, University of Queensland, St Lucia QLD 4072, Australia.
e-mail: ethan.scott@uq.edu.au

1. Marques, J. C., Li, M., Schaak, D., Robson, D. N. & Li, J. M. *Nature* **577**, 239–243 (2020).
2. Ahrens, M. B. *et al. Nature* **485**, 471–477 (2012).
3. Vanwalleghem, G. C., Ahrens, M. B. & Scott, E. K. *Curr. Opin. Neurobiol.* **50**, 136–145 (2018).
4. Kim, D. H. *et al. Nature Methods* **14**, 1107–1114 (2017).
5. Cohen, J. D., McClure, S. M. & Yu, A. J. *Phil. Trans. R. Soc. Lond. B* **362**, 933–942 (2007).
6. Flavell, S. W. *et al. Cell* **154**, 1023–1035 (2013).
7. Kunst, M. *et al. Neuron* **103**, 21–38 (2019).
8. Hildebrand, D. G. C. *et al. Nature* **545**, 345–349 (2017).

This article was published online on 18 December 2019.

Astronomy

Not all fast radio bursts are created equal

Matthew Bailes

Astronomical signals called fast radio bursts remain enigmatic, but a key discovery has now been made. A second repeating fast radio burst has been traced to its host galaxy, and its home bears little resemblance to that of the first. **See p.190**

Millisecond-duration, extragalactic radio flashes called fast radio bursts[1,2] (FRBs) present many puzzles that are strikingly similar to past mysteries concerning phenomena now known as quasars and γ-ray bursts. Like FRBs, these phenomena seemed to be uniformly distributed in the sky, but of almost impossible luminosity if cosmological in origin. Ultimately, it was revealed that quasars are associated with supermassive black holes, and γ-ray bursts with the formation of black holes after supernovae (stellar explosions) or with mergers of stellar remnants called neutron stars. Repeating FRBs provide the opportunity to identify the homes and possible progenitors of FRB sources. The first known repeater was localized to a star-forming region in a tiny, metal-poor galaxy[3]. On page 190, Marcote *et al.*[4] pinpoint the home of a second repeater and find it to be very different.

As an FRB travels through the ionized intergalactic medium, its radio waves interact with free electrons and are slowed. High-frequency waves are less affected than low-frequency ones, so the former arrive at an FRB detector slightly sooner than the latter. This time difference is proportional to the total number of electrons between the FRB source and the detector, and can therefore be used to estimate the enormous distance to the source. Early FRB detectors could not unambiguously identify an FRB host galaxy because their localization resolutions spanned hundreds or even thousands of potential host galaxies. But now, multi-element radio telescopes such as the Australian Square Kilometre Array Pathfinder (ASKAP) and the Deep Synoptic Array ten-antenna prototype (DSA-10) in California have high enough resolution that host galaxies can be easily identified.

"The host galaxy is about 200 times farther from Earth than is our nearest neighbour, the Andromeda galaxy."

At the same time, a radio telescope known as the Canadian Hydrogen Intensity Mapping Experiment[5] (CHIME) is revealing that many FRBs repeat. The individual bursts of a repeater tend to be clustered in time, which means that the probability of catching a burst is higher when a repeater is active. Because CHIME surveys the sky daily, it is in an ideal position to trigger other telescopes to carry

© 2020 Springer Nature Limited. All rights reserved.

https://doi.org/10.1038/d41586-019-03894-6

News & views

Astronomy

Not all fast radio bursts are created equal

Matthew Bailes

Astronomical signals called fast radio bursts remain enigmatic, but a key discovery has now been made. A second repeating fast radio burst has been traced to its host galaxy, and its home bears little resemblance to that of the first.

Millisecond-duration, extragalactic radio flashes called fast radio bursts[1,2] (FRBs) present many puzzles that are strikingly similar to past mysteries concerning phenomena now known as quasars and γ-ray bursts. Like FRBs, these phenomena seemed to be uniformly distributed in the sky, but of almost impossible luminosity if cosmological in origin. Ultimately, it was revealed that quasars are associated with supermassive black holes, and γ-ray bursts with the formation of black holes after supernovae (stellar explosions) or with mergers of stellar remnants called neutron stars. Repeating FRBs provide the opportunity to identify the homes and possible progenitors of FRB sources. The first known repeater was localized to a star-forming region in a tiny, metal-poor galaxy[3]. Writing in *Nature*, Marcote *et al.*[4] pinpoint the home of a second repeater and find it to be very different.

As an FRB travels through the ionized intergalactic medium, its radio waves interact with free electrons and are slowed. High-frequency waves are less affected than low-frequency ones, so the former arrive at an FRB detector slightly sooner than the latter. This time difference is proportional to the total number of electrons between the FRB source and the detector, and can therefore be used to estimate the enormous distance to the source. Early FRB detectors could not unambiguously identify an FRB host galaxy because their localization resolutions spanned hundreds or even thousands of potential host galaxies. But now, multi-element radio telescopes such as the Australian Square Kilometre Array Pathfinder (ASKAP) and the Deep Synoptic Array ten-antenna prototype (DSA-10) in California have high enough resolution that host galaxies can be easily identified.

At the same time, a radio telescope known as the Canadian Hydrogen Intensity Mapping Experiment[5] (CHIME) is revealing that many FRBs repeat. The individual bursts of a repeater tend to be clustered in time, which means that the probability of catching a burst is higher when a repeater is active. Because CHIME surveys the sky daily, it is in an ideal position to trigger other telescopes to carry out localizations. Marcote and colleagues caught four bursts of one of the repeaters discovered by CHIME and used these observations to determine the precise location of the repeater in a nearby galaxy (Fig. 1). This galaxy is about 200 times farther from Earth than is our nearest neighbour, the Andromeda galaxy.

The first known repeater differed in at least three ways from many of the one-off FRBs. First, it showed peculiar changes in intensity as a function of frequency and time, akin to intense solar radio emissions called solar radio bursts. Second, it traversed a highly magnetized plasma that led to a twist in its polarization as a function of frequency. And third, it was emitted near a supermassive black hole. Although one-off FRBs are harder to study than are repeaters, they are known to sometimes have complicated, multi-component temporal profiles[6], similar to the emissions from magnetars (highly magnetized neutron stars).

Could it be that one-off FRBs and repeaters are like short- and long-duration γ-ray bursts — two different classes of phenomenon that just happen to have a similar emission timescale and luminosity? Or is there a theory of FRBs that would link the two classes? Most FRB theories involve a compact astronomical object, and emissions that are associated with extremely powerful magnetic fields or rapidly spinning (and therefore energetic) neutron stars. Some scientists, desperate for an explanation, have even proposed alien transmissions[7].

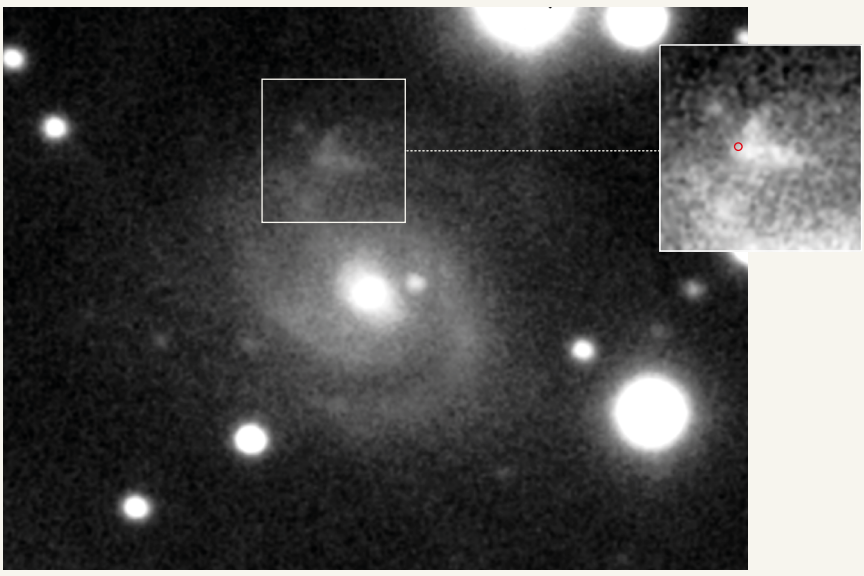


Figure 1 | Localization of a repeating fast radio burst. Fast radio bursts (FRBs) are bright, extragalactic radio flashes of unknown physical origin. Marcote *et al.*[4] have determined the precise location of a source of repeating FRBs (indicated by the red circle). Unlike the previous localization of a repeating FRB[3], this source was found in a fairly ordinary, massive spiral galaxy. The image is taken using optical light, and a higher-contrast zoom-in of the relevant region of the galaxy is included. (Image from ref. 4.)

© 2020 Springer Nature Limited. All rights reserved.

It is possible to envisage a unifying model in which FRB sources have an energy reservoir (whether magnetic, kinetic or gravitational) that is plentiful when the source is produced near a star-forming region, causing frequent bursts. As this reservoir becomes depleted, the bursts are separated by increasingly long gaps, until they cease altogether. Such behaviour is similar to that of 'glitches' (sudden changes in spin frequency) in a type of neutron star called a pulsar. Young pulsars often glitch prolifically, but older pulsars can go for decades without exhibiting a glitch. Frequently repeating FRBs would be easier to find than those that have a lower repetition rate, but verifying that they are repeaters would become increasingly difficult as their repetition timescales approached the lifetime of FRB experiments (currently years).

The first known repeater seemed to fit such a model. It could have been a young magnetar emitting from a relatively exotic location — near a supermassive black hole in the star-forming region of a small, metal-poor galaxy. Such an environment is very different from that of most young neutron stars in our own Galaxy. However, Marcote and colleagues found that the source of their repeater is not particularly magnetic, is nowhere near a supermassive black hole and is in a fairly normal galaxy. Even so, the source is close to a curious, V-shaped region of stars that only adds to the mystery of FRB origins (Fig. 1).

In the past year, ASKAP has localized several one-off FRBs[8,9] and found them to inhabit a wide range of galaxy types. The beauty of the ASKAP results and those of Marcote *et al.* lies in the accuracy and reliability of the localizations. These features are not available to relatively compact telescopes[10], such as DSA-10, which can only hint at the identity of the relevant galaxy (think of this as the FRB's home city), rather than pinpointing the exact environment within it (the FRB's home address).

Ultimately, we will discover that, for each class of FRB, either these signals are produced near active sites of star formation or there is no particular correlation. If they originate near the sites of supernovae, the FRBs probably arise from young, highly magnetic neutron stars. If their origins are farther from such sites, they might be produced by time bombs of gravitational energy (such as merging neutron stars) and have diverse spatial distributions similar to those of short-duration γ-ray bursts, which are now known to be associated with neutron-star coalescence. Curiously, because the luminosity of a typical quasar is not dissimilar to the peak luminosity of an FRB, we cannot yet completely discount the involvement of supermassive black holes. But Marcote and colleagues' results argue against that possibility, and the timescales for quasar variability are more likely to be days to months than milliseconds.

Rates of FRB discovery and host-galaxy identification are likely to greatly increase as more detectors come online. Results from CHIME are widely anticipated, and facilities such as the MeerKAT radio telescope in South Africa and the Five-hundred-meter Aperture Spherical radio Telescope (FAST) in China will soon come online. Do not be surprised if there is more than one way to generate FRBs — without the need for aliens.

Matthew Bailes is at the Centre for Astrophysics & Supercomputing, Swinburne University of Technology, Hawthorn, Victoria 3122, Australia.
e-mail: mbailes@swin.edu.au

1. Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J. & Crawford, F. *Science* **318**, 777–780 (2007).
2. Thornton, D. *et al. Science* **341**, 53–56 (2013).
3. Chatterjee, S. *et al. Nature* **541**, 58–61 (2017).
4. Marcote, B. *et al. Nature* https://doi.org/10.1038/s41586-019-1866-z (2020).
5. The CHIME/FRB Collaboration *et al. Astrophys. J.* **885**, L24 (2019).
6. Farah, W. *et al. Mon. Not. R. Astron. Soc.* **478**, 1209–1217 (2018).
7. Lingam, M. & Loeb, A. *Astrophys. J.* **837**, L23 (2017).
8. Bannister, K. W. *et al. Science* **365**, 565–570 (2019).
9. Prochaska, J. X. *et al. Science* **366**, 231–234 (2019).
10. Ravi, V. *et al. Nature* **572**, 352–354 (2019).

© 2020 Springer Nature Limited. All rights reserved.

Review

A brief history of human disease genetics

https://doi.org/10.1038/s41586-019-1879-7

Received: 16 July 2019

Accepted: 13 November 2019

Published online: 8 January 2020

Melina Claussnitzer[1,2,3], Judy H. Cho[4,5,6], Rory Collins[7,8], Nancy J. Cox[9], Emmanouil T. Dermitzakis[10,11], Matthew E. Hurles[12], Sekar Kathiresan[2,13,14], Eimear E. Kenny[4,6,15], Cecilia M. Lindgren[2,16,17], Daniel G. MacArthur[2,13,18], Kathryn N. North[19,20], Sharon E. Plon[21,22], Heidi L. Rehm[2,13,18,23], Neil Risch[24], Charles N. Rotimi[25], Jay Shendure[26,27,28], Nicole Soranzo[12,29] & Mark I. McCarthy[17,30,31,32]*

A primary goal of human genetics is to identify DNA sequence variants that influence biomedical traits, particularly those related to the onset and progression of human disease. Over the past 25 years, progress in realizing this objective has been transformed by advances in technology, foundational genomic resources and analytical tools, and by access to vast amounts of genotype and phenotype data. Genetic discoveries have substantially improved our understanding of the mechanisms responsible for many rare and common diseases and driven development of novel preventative and therapeutic strategies. Medical innovation will increasingly focus on delivering care tailored to individual patterns of genetic predisposition.



Anniversary collection: go.nature.com/nature150

For almost all human diseases, individual susceptibility is, to some degree, influenced by genetic variation. Consequently, characterizing the relationship between sequence variation and disease predisposition provides a powerful tool for identifying processes fundamental to disease pathogenesis and highlighting novel strategies for prevention and treatment.

Over the past 25 years, advances in technology and analytical approaches, often building on major community projects—such as those that generated the human genome sequence[1] and elaborated on that reference to capture sites of genetic variation[2–6]—have enabled many of the genes and variants that are causal for rare diseases to be identified and enabled a systematic dissection of the genetic basis of common multifactorial traits. There is growing momentum behind the application of this knowledge to drive innovation in clinical care, most obviously through developments in precision medicine. Genomic medicine, which was previously restricted to a few specific clinical indications, is poised to go mainstream.

This Review charts recent milestones in the history of human disease genetics and provides an opportunity to reflect on lessons learned by the human genetics community. We focus first on the long-standing division between genetic discovery efforts targeting rare variants with large effects and those seeking alleles that influence predisposition to common diseases. We describe how this division, with its echoes of the century-old debate between Mendelian and biometric views of human genetics, has obscured the continuous spectrum of disease risk alleles—across the range of frequencies and effect sizes—observed in the population, and outline how genome-wide analyses in large biobanks are transforming genetic research by enabling a comprehensive perspective on genotype–phenotype relationships. We describe how the expansion in the scale and scope of strategies for enumerating the functional consequences of genetic variation is transforming the torrent of genetic discoveries from the past decade into mechanistic insights, and the ways in which this knowledge increasingly underpins advances in clinical care. Finally, we reflect on some of the challenges and opportunities that confront the field, and the principles that will, over the coming decade, drive the application of human genetics to enhance understanding of health and disease and maximize clinical benefit.

[1]Beth Israel Deaconess Medical Center, Harvard Medical School, Boston, MA, USA. [2]Broad Institute of MIT and Harvard Cambridge, Cambridge, MA, USA. [3]Institute of Nutritional Science, University of Hohenheim, Stuttgart, Germany. [4]Charles Bronfman Institute for Personalized Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [5]Department of Medicine, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [6]Department of Genetics and Genomic Sciences, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [7]Nuffield Department of Population Health (NDPH), University of Oxford, Oxford, UK. [8]UK Biobank, Stockport, UK. [9]Vanderbilt Genetics Institute and Division of Genetic Medicine, Department of Medicine, Vanderbilt University Medical Center, Nashville, TN, USA. [10]Department of Genetic Medicine and Development, University of Geneva Medical School, Geneva, Switzerland. [11]Health 2030 Genome Center, Geneva, Switzerland. [12]Wellcome Sanger Institute, Hinxton, UK. [13]Center for Genomic Medicine, Massachusetts General Hospital, Boston, MA, USA. [14]Verve Therapeutics, Cambridge, MA, USA. [15]Center for Genomic Health, Icahn School of Medicine at Mount Sinai, New York, NY, USA. [16]Big Data Institute at the Li Ka Shing Centre for Health Information and Discovery, University of Oxford, Oxford, UK. [17]Wellcome Centre for Human Genetics, Nuffield Department of Medicine, University of Oxford, Oxford, UK. [18]Analytic and Translational Genetics Unit, Massachusetts General Hospital, Boston, MA, USA. [19]Murdoch Children's Research Institute, Parkville, Victoria, Australia. [20]University of Melbourne, Parkville, Victoria, Australia. [21]Departments of Pediatrics and Molecular and Human Genetics, Baylor College of Medicine, Houston, TX, USA. [22]Texas Children's Cancer Center, Texas Children's Hospital, Houston, TX, USA. [23]Department of Pathology, Harvard Medical School, Boston, MA, USA. [24]Institute for Human Genetics, University of California San Francisco, San Francisco, CA, USA. [25]Center for Research on Genomics and Global Health, National Human Genome Research Institute, Bethesda, MD, USA. [26]Department of Genome Sciences, University of Washington, Seattle, WA, USA. [27]Brotman Baty Institute for Precision Medicine, Magnuson Health Sciences Building, Seattle, WA, USA. [28]Howard Hughes Medical Institute, Seattle, WA, USA. [29]Department of Haematology, University of Cambridge, Cambridge, UK. [30]Oxford Centre for Diabetes, Endocrinology and Metabolism, Oxford, UK. [31]Oxford NIHR Biomedical Research Centre, Oxford University Hospitals NHS Foundation Trust, John Radcliffe Hospital, Oxford, UK. [32]Human Genetics, Genentech, South San Francisco, CA, USA. *e-mail: mccarthy.mark@gene.com

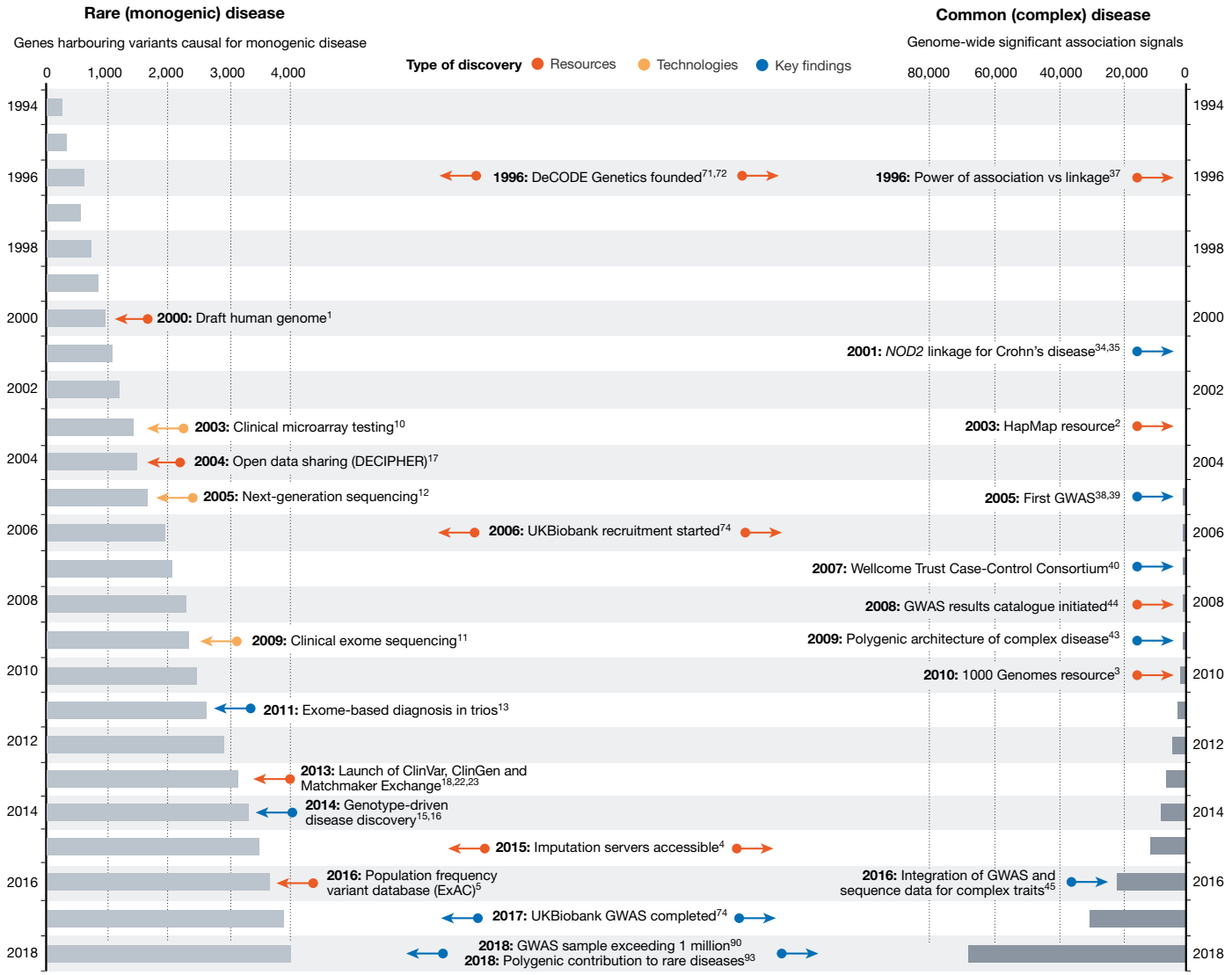


Fig. 1 | Growth in the discovery of disease-associated genetic variation. The cumulative numbers of genes harbouring variants causal for rare, monogenic diseases and traits and of significant GWAS associations implicated in common, complex diseases and traits are shown. Left, the advent of high-throughput sequencing technologies and availability of reference genomes from diverse populations has supported a fourfold increase in the discovery of rare disease-causing genes between 1999 and 2019. Right, international efforts such as the Human Genome Project and the HapMap Project, combined with GWAS and sequencing studies, have supported identification of more than 60,000 genetic associations across thousands of human diseases and traits. Centre, more recent developments have brought a synthesis of the rare- and common-variant approaches based around the combination of sequence-informed analyses in large cohorts. Key events contributing to these themes are depicted in the timeline. GA4GH, Global Alliance for Genomics and Health[160]; ExAC, Exome Aggregation Consortium[5].

Rare diseases, rare variants

During the 1980s and 1990s, efforts to map disease genes were focused on rare, monogenic and syndromic diseases and were mostly driven by linkage analysis and fine mapping within large multiplex pedigrees. Localization of genetic signals was typically followed by Sanger sequencing of the genes found to map within the linked locus to identify disease-causing alleles. Assessments of pathogenicity, based on segregation of a putatively causal variant with disease across multiple families and evidence that the risk genotype was absent in healthy individuals, were typically followed by confirmatory functional studies in cellular and animal models. This path to gene identification was laborious; nevertheless, by 2000, around 1,000 of the estimated 7,000 single-gene inherited diseases had been characterized, including many with substantial biomedical impact, such as Huntington's disease and cystic fibrosis[7–9].

Completion of the draft human genome sequence[1] reduced many of the obstacles to disease-gene mapping and propelled a fourfold increase in the genes implicated as causal for rare, single-gene disorders (Fig. 1). Microarray-based detection of structural variation[10] and exome- and genome-wide sequencing[11,12] have been pivotal, bolstered by in silico analysis and prioritization of the discovered genetic variants. Increasing availability of reference datasets cataloguing population genetic variation across diverse ethnic backgrounds has supported robust causal inference[2,3,5,6]. More recently, the adoption of high-throughput sequencing technologies has enabled the full range of causal genetic variation, from single mutations to large structural rearrangements, to be identified in a single assay. These technologies have extended from research into clinical usage, driving earlier and faster diagnosis for genetic disorders.

Reduced reliance on multiplex pedigrees in favour of collections of affected cases, often with parents[13], has proven decisive in identifying

new dominant disorders, many of which were previously considered recessive[14]. Increasingly, discovery of rare disease genes has transitioned from genetic characterization of small numbers of individuals with similar clinical presentations to genome-wide sequencing of larger cohorts of phenotypically diverse patients. This genotype-driven approach has revealed new disorders associated with more variable clinical presentation[15,16].

A more systematic approach to data sharing has been critical, both for the characterization of new disorders and diagnostic interpretation of potential causal alleles. The value of sharing genetic and phenotypic data from those thought to harbour rare undiagnosed genetic diseases has fostered global collaborative networks (for example, Matchmaker Exchange, DECIPHER and GeneMatcher) designed to match patients with similar genetic variants and/or phenotypic manifestations, even across continents[17–19]. Interactions between researchers and families with rare disease have enabled natural history studies to be driven by family support groups positioned to initiate data collection from patient cohorts once a causal gene is discovered[20].

Clinical translation of these technologies has benefited from a series of information resources, including open databases of genes associated with rare disorders (for example, OMIM and ORPHANET)[21], clinically interpreted variants (for example, ClinVar and ClinGen)[22,23] and patient records (for example, DECIPHER and MyGene2 (https://mygene2.org/MyGene2))[17]. Access to resources that catalogue genetic variation across populations (such as ExAC and its successor gnomAD)[5,6] has enabled the confident exclusion of genetic variants too common in population-level data to be plausible causes of rare, penetrant early-onset genetic diseases[24]. These analyses have reduced the contamination of databases with variants erroneously interpreted as causal for disease, and are addressing the overestimation of disease penetrance arising from the historical focus on multiplex pedigrees[25]. Improved recognition of the variable penetrance of many 'monogenic' disease alleles has invigorated efforts to identify the genetic and environmental modifiers responsible[26,27].

Although huge strides have been made in associating specific genes with particular disorders, establishing the causal role of individual variants within those genes remains problematic, and many patients with suspected rare genetic diseases are left without a definitive diagnosis[28]. Even for variants with established causality, the penetrance is often unclear. Resolving these uncertainties represents the central challenge for the field. Aggregation of sequencing data from large numbers of affected cases and population reference samples will provide the evidence base required for robust interpretation of variants. Highly parallelized in vitro cellular assays that allow assessment of the functional effects of all variants in a disease-associated gene can transform interpretation of novel variants[29], although developing well-calibrated functional assays predictive of pathogenicity for all disease genes represents a daunting prospect. Direct functional genomic exploration of accessible and disease-relevant tissues from patients using RNA sequencing and DNA methylation assays[30,31] can identify previously cryptic causal genetic variants, particularly in under-explored regions outside protein-coding genes[32,33]. Developments in each of these areas will extend the range of variants and genes for which diagnostic and prognostic clinical information can be provided to patients and their families.

Common diseases, common variants

Efforts to apply the approach—linkage analysis in multiplex pedigrees—that had been so successful for the high-penetrance variants responsible for Mendelian disease were, with notable exceptions[34–36], largely unsuccessful for common, later-onset traits with more complex multifactorial aetiologies, such as asthma, diabetes and depression. Recognition that association-based methods, focused on detecting phenotype-related differences in variant allele frequencies might have greater traction for identifying less penetrant common alleles redirected attention to analysis of case–control samples[37]. However, initial efforts targeting variants within 'candidate' genes were plagued by inadequate power, unduly liberal thresholds for declaring significance and scant attention to sources of bias and confounding, resulting in overblown claims and failed replication.

Systematic efforts to characterize genome-wide patterns of genomic variation, initially through the HapMap Consortium[2], proved catalytic, demonstrating that the allelic structure of the genome was segmented into haplotype blocks, each containing sets of correlated variants. Recognition that this configuration could support genome-wide surveys of association energized the technological innovation—in the form of massively parallel genotyping arrays—to make such studies possible (Fig. 1). Early wins in acute macular degeneration[38] and inflammatory bowel disease[39] were encouraging, and progress on several fronts—expansion of study size, denser genotyping arrays, novel strategies for imputation, attention to biases and appropriate significance thresholds—delivered robust associations across a range of diseases[40]. Most variants uncovered by these early genome-wide association studies (GWAS) were common, with more subtle effects than many had anticipated. A host of trait-specific consortia formed, covering diverse dichotomous and quantitative phenotypes, to accelerate genetic discovery through the aggregation and meta-analysis of data from multiple GWAS studies[41–43]. Many tens of thousands of robust associations were identified[44]. Recently, increased access to exome and whole-genome sequence data has, through both direct association analysis[45,46] and imputation[3,4], extended discovery to low-frequency and rare alleles previously inaccessible to GWAS.

In the decade since the first GWAS, understanding of the genetic basis of common human disease has been transformed. The disparity between the observed effects of the variants first identified by GWAS and estimates of overall trait heritability (the 'missing heritability' conundrum) is now largely resolved[47]. Common diseases are not simply aggregations of related Mendelian conditions: for most complex traits, genetic predisposition is shared across thousands of mostly common variants with individually modest effects on population risk[41,43].

Although the collective contribution of low-frequency and rare risk alleles to overall trait variability appears modest compared with that attributable to common variants[45,48], the rare risk alleles detected in current sample sizes necessarily have large phenotypic effects and are proportionately more likely to be coding, enhancing their value for biological inference. Founder populations (such as those from Finland and Iceland) have provided multiple examples of otherwise rare risk alleles driven to higher frequency locally through drift and/or selection[49–52]. In addition, studies in populations with high rates of consanguinity make it possible to identify individuals homozygous for otherwise rare loss-of-function alleles, the basis for a 'human knockout' project to systematically investigate the phenotypic consequences of gene disruption in humans[53,54].

For most diseases, large-scale GWAS-aggregation efforts have been disproportionately powered by information from individuals of European descent[55]. Whereas patterns of genetic predisposition appear broadly similar across major population groups and many common risk alleles discovered in one population group are detectable in others, allele frequencies can vary substantially; extending GWAS and sequencing studies to diverse populations will surely generate a rich harvest of novel risk alleles.

The relative contributions of common and rare variants indicate that, for many traits, particularly those with post-reproductive onset, purifying selection has had only limited effect[45,56]. For a few risk alleles, hallmarks of balancing selection reflect increased carrier survival, usually through protection from infectious diseases. This includes well-known examples of alleles maintained at high frequency in populations of African descent[57,58].

While the extensive linkage disequilibrium within human populations has been essential to discovery in GWAS, high correlation between adjacent variants frustrates mapping of the specific variants responsible for these associations. Increasing sample size, improved access to trans-ethnic data, and more representative imputation reference panels[3] provide a path to improved resolution of the causal variants[59] and clues to the molecular mechanisms through which they operate. Functional interpretation is easiest for causal variants within coding sequences; however, most common disease-risk variants map to noncoding sequences, and are presumed to influence predisposition through effects on transcriptional regulation. In these cases, mechanistic inference depends on connecting association signals to their downstream targets (see below). For many traits, there is clear convergence between common-variant association signals and genes implicated in monogenic forms of the same disease, as well as enrichment of GWAS signals in regulatory elements specifically active in cell types consistent with known disease biology[60,61]. This provides reassurance that, even as the number of association signals for a given disease proliferates, the genetic associations uncovered will coalesce around molecular and cellular processes with a core role in pathogenesis[62,63].

Importantly, the signals discovered by GWAS have revealed many unexpected insights into the biological basis of complex disease. Examples include the role of complement in the pathogenesis of acute macular degeneration[38], synaptic pruning in schizophrenia[64] and autophagy in inflammatory bowel disease[65]. In addition, as inherited sequence variation is a prominent cause of phenotypic variation (but the reverse is not true), risk variants identified by GWAS have value as genetic instruments, mapping causal relationships between traits and inferring contributions made by circulating biomarkers and environmental exposures to disease development[66].

As described below, findings from GWAS have increasing translational impact through identification of novel therapeutic targets[67], prioritization (and deprioritization) of existing ones[68] and development of polygenic scores that quantify individual genetic risk[69].

Comprehensive genotype–phenotype maps

The historical division of disease-gene discovery into monogenic and polygenic strands arose from development and implementation of analytical approaches—family-based linkage and case–control association[37]—that are best-suited for detecting particular subsets of causal alleles. This obscured the true state of nature, with disease-risk alleles being distributed across a continuous spectrum of frequencies and effect sizes. In addition, the trait- and disease-specific perspective of early GWAS discovery (mostly reliant on case–control studies) was poorly equipped to investigate the contribution of genetic variants to phenotypic effects that are nested within or spread across classical disease definitions. Recent developments have enabled a more holistic perspective on genotype–phenotype relationships (Fig. 1).

One major advance has been the increasing availability of large prospective population-based cohorts. These biobank efforts, pioneered in studies such as the Framingham Cohort[70] and the efforts of DeCODE in Iceland[71,72], now encompass a growing inventory of national cohorts in North America, Europe, Asia and beyond[73–76]. The UK Biobank study, including 500,000 largely healthy, middle-aged participants has been particularly influential, transforming human genetic research in part through permissive data-sharing policies that have allowed multiple research groups to analyse the data[74]. Efforts to make clinical data embedded in electronic health records and registries available for research[77,78] mean that biobanks increasingly provide access to a wide range of demographic, clinical and lifestyle data, captured in harmonized, systematic fashion from large, often multi-ethnic collections of individuals. For millions of biobank participants, this rich phenotypic information has been combined with genome-wide genetic data. There are nascent efforts to capture transcriptomic, proteomic and metabolomic phenotypes, although these are not yet at equivalent scale to the genetic data[79,80]. Biobank analyses have provided more generalizable estimates of the relevance of genetic risk factors in the context of the separate and joint effects of non-genetic factors[81]. Increasingly, integration with healthcare data brings a longitudinal dimension to phenotypic characterization, which facilitates analyses of disease progression and lifelong disease risk[82].

The rich phenotypic scope of these cohorts has enabled variants of interest to be interrogated for associations across the gamut of available phenotypes. These phenome-wide association studies (PheWAS) have revealed the extent to which many variants have pleiotropic effects across multiple traits[83]. Some of these relationships are expected, such as the impact of obesity variants on risk of hepatic steatosis and type 2 diabetes[84] or variants that influence multiple autoimmune conditions[85]. Others connect diseases and traits in surprising ways, highlighting shared polygenic, pleiotropic effects and cell-type specificity, and delivering insights into shared biology and overlapping mechanisms[86,87]. These findings inform the prioritization of therapeutic targets, providing clues to potential on-target side effects and opportunities for drug repurposing[87–89].

The second enabler of inclusive, systematic analysis of genotype–phenotype relationships has been access to whole-genome sequence data. The scale of genetic analysis based on sequence data still lags behind that of genome-wide genotyping data (the largest sequence-based datasets are one tenth the size of the largest GWAS[90–92]), although reductions in sequencing costs are decreasing the differential. Most direct analysis of high-throughput sequence data has focused on the coding regions. Strategies for assigning variant function and jointly analysing sets of variants of similar functional effect have enabled aggregate, gene-level tests of rare functional-variant association that are often better powered than single-variant tests[91,92]. However, the principal benefit to date of whole-genome sequence data to genetic discovery has been to bolster array-based access to lower-frequency alleles, either directly, through their inclusion on genotyping platforms, or indirectly, through imputation from sequence-based reference samples[3,4].

These developments have enabled researchers to bridge the gap between the monogenic and polygenic realms, identifying common variant modifiers of monogenic phenotypes contributing to the variable expression of rare, large-effect alleles[26,93], and low-frequency and rare variants that influence common multifactorial traits[94,95]. This enables more rigorous evaluation of the contribution of rare and common variants to trait susceptibility[48] and supports the enumeration of 'allelic series' (sets of alleles of varying frequency, effect size and direction that disrupt the same gene) critical for studies of disease mechanism and therapeutic target optimization[89,96]. These developments are rapidly converging towards the ultimate destination: a comprehensive matrix of the effect of all observable genetic variants across the widest possible range of cross-sectional and longitudinal biomedical phenotypes. Success in this endeavour depends on ever greater harmonization between, and integration of results from, individual studies through sustained investments in data sharing.

Adding function

From the first linkage maps to whole-genome sequencing of large cohorts, human genetics has deployed increasingly sophisticated and inherently systematic approaches for mapping the genetic factors that underlie traits and diseases. However, progress in determining how these variants influence disease, through systematic interrogation of their functional effects on molecular, cellular and physiological processes, has been far slower.

For monogenic diseases, for which the alleles responsible are typically rare, penetrant and coding, genetic approaches have generally been both necessary and sufficient to implicate a gene as causal[28].

However, as efforts to elucidate the genetic basis of Mendelian disorders progress towards completion[97], functional studies remain important to understand the mechanisms by which disruptive variation within a causal gene leads to disease phenotypes. Unlike common diseases, the clarity of causation for Mendelian disorders usually simplifies the task of generating models (including human cells and organoids or rodents) to connect genotype to organismal phenotype; these have led to many critical insights into the biology of health and disease in humans[98,99]. In addition, for genes harbouring variants with medically actionable consequences (as with the *BRCA1* and *BRCA2* mutations that are causal for early-onset breast and ovarian cancer), functional studies can support the translational interpretation of novel alleles identified by medical sequencing[29].

For common diseases, functional studies have a more fundamental role. Although tens of thousands of associations have been discovered across thousands of common human diseases and traits[44], multiple factors have frustrated efforts to convert these genetic signals to knowledge about causal variants, genes and mechanisms. For the common variants that underlie the bulk of complex-disease risk, the resolution of association mapping is often limited by the haplotype structure of the human genome[2–4]. Furthermore, most GWAS associations map to the noncoding genome and thus lack a direct address to the gene that mediates their effects. Growing appreciation of the pervasive role of pleiotropy complicates matters: many variants identified by GWAS are associated with multiple traits and exert diverse effects across multiple cell types[100].

To date, relatively few studies have achieved the goal of connecting variants causal for complex traits to the molecular and cellular functions that mediate that predisposition. One early success described how regulatory variants that modulate *SORT1* expression influence low-density lipoprotein cholesterol and myocardial infarction risk[101]. More recent examples have focused on the relationship between obesity-associated variants intronic to *FTO*, altered expression of *IRX3* and *IRX5*, and adipocyte[102] and hypothalamic[103] function. Similar functional descriptions have been reported for individual loci implicated in schizophrenia[64], cardiovascular disease[104], type 2 diabetes[105] and Alzheimer's disease[106], among others.

Over the past decade, the challenge for the functional genomics community has been to convert this 'one-locus-at-a-time' workflow to a systematic, multidimensional, integrative approach able to deliver genome-scale functional analyses to match genome-wide variant discovery (Fig. 2). At the molecular level, one cornerstone has been generation of genome-wide catalogues of functional activity. For example, the ENCODE and Roadmap Epigenomics projects have generated maps of histone modifications, transcription-factor binding, chromatin accessibility, three-dimensional genome structure and other regulatory annotations across hundreds of cell types and tissues [107,108]. The patterns of genomic overlap between these data and GWAS results enable the functional inference of risk variants, deliver clues to the specific cell types driving disease pathogenesis[60,109] and accelerate locus-specific mechanistic insights.

In parallel, there has been a scaling of efforts to connect trait-associated regulatory variants to the genes and processes that they regulate in cell types relevant to the disease of interest[110,111]. For example, the GTEx (Genotype-Tissue Expression) consortium has mapped thousands of expression quantitative trait loci (QTLs) across hundreds of individuals and dozens of tissues[112]. Further clues to the relationships between regulatory variants and their effector genes can be gathered from DNA proximity assays (such as Hi-C) and single-cell data[113] (Fig. 2). Programs such as HubMAP[114] and the Human Cell Atlas[115] are set to deliver comprehensive, high-resolution reference maps of individual human cell types across diverse developmental stages, providing new opportunities to understand how regulatory genetic variation results in cellular and organismal phenotypes.

Efforts to probe the clinical consequences of coding alleles with large phenotypic effects (particularly null alleles) in humans[53,54] and across diverse animal models[116] represent powerful strategies for extending functional analyses to the whole-body level. Connections between genetic variation and circulating proteomic and metabolomic data provide additional mechanistic links between cellular events and whole-body physiology[79,80]. These efforts are paralleled by PheWAS approaches[83], which, by mapping variant effects across the range of traits available in biobanks and EMRs, can inform priors for cell types and pathways at individual loci. Importantly, whereas early studies typically linked GWAS risk alleles to data from a single functional assay, the focus is increasingly on maximizing biological insight through the multi-dimensional integration of multiple genome-wide data types using approaches such as heritability partitioning[117], functional enrichment analyses[60,109], integration of the three-dimensional genome structure[118] and deep convolutional neural networks[119,120].

Although QTL analyses can implicate a haplotype in a molecular, cellular or organismal phenotype, they are, in isolation, insufficient to define the specific causal variants responsible. To address this, there has been rapid maturation of technologies, such as massively parallel reporter assays[121–123] and CRISPR genome editing, to support functional characterization of targeted sequence perturbations at scale. Variations on these methods enable the functional evaluation of genes (via knockout screens[124]), regulatory elements (using CRISPR interference and CRISPR activation screens[125,126]), and genetic variants (base editors[127]) at increasing scale and resolution[29]. Combined with complex readouts—including high-content imaging[128] and single-cell transcriptomics and epigenomics[129,130]—these methods can generate empirical 'truth' data, supporting the development of in silico models to predict causal variants, effector transcripts[126] and cellular effects. In due course, such models should reduce the need for exhaustive experimental characterization of function for all variants across all cell types.

The goal of such efforts is to enumerate the cascade of molecular events that underlie observed genotype–phenotype associations using physiologically relevant cellular systems (from primary cells to organoids and 'organ-on-chip' designs) and whole-body assays appropriate to the disease of interest. Collectively, strategies that offer large-scale functional evaluation of variants and genes of interest will reduce (but probably not eliminate) the intensive effort required for 'final mile' validation of disease mechanisms in dedicated systems, thereby accelerating downstream translational application.

Clinical implementation

Medical genetics, as applied to rare diseases, has been characterized by the rapid application in the clinic of the transformative genomic technologies that drove initial research discoveries. There are now targeted genetic tests for nearly all clinical presentations attributable to large-impact alleles, alongside more extensive genome-sequencing assays that, when necessary, enable interrogation of a longer list of relevant genes. Genetic testing for symptomatic individuals and at-risk relatives occurs routinely in many medical specialties. In parallel, the use of somatic cancer testing has increased as therapies targeted to specific mutational events have entered clinical practice (these developments are reviewed elsewhere[131,132]).

For patients with symptoms that indicate a probable monogenic aetiology (such as retinal degeneration, hearing loss or cardiomyopathy), targeted panels are typically the platform of choice[133], although they are increasingly performed on a more extensive sequence backbone. For more complex phenotypes—those without a clear match to a specific syndrome, such as neurodevelopmental disorders and multiple congenital anomalies—testing has gravitated towards early deployment of exome and genome-sequencing platforms that offer speedy resolution of what has historically often been a traumatic diagnostic odyssey[15,134]. The power of genomic diagnosis is especially clear for those presenting with monogenic neurodevelopmental disorders and critically ill infants[135,136]. Sequencing of the parent–offspring trio can

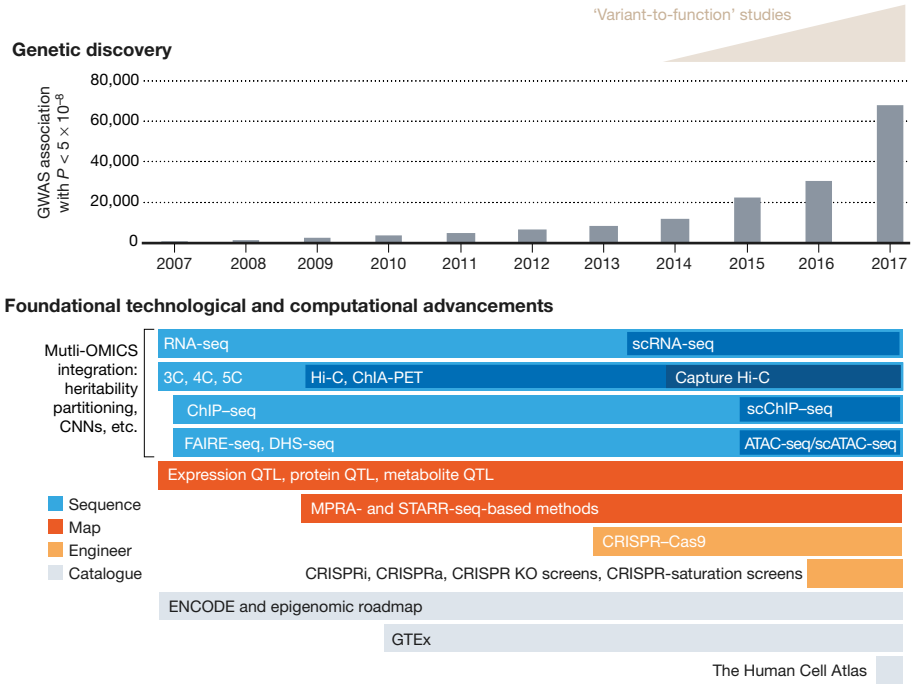


Fig. 2 | Genetic discovery is paralleled by advances in functional genomics technologies. Top, the growth in the number of genetic loci associated by GWAS with human traits and diseases (bars) and of variant-to-function studies (area under line, not to scale). Bottom, foundational technological and computational advances over the last decade that enabled (1) development of systematic, genome-wide catalogues of functional elements across multiple cell types and tissues (blue); (2) mapping of QTLs in the context of gene expression, metabolites, proteins and regulatory elements (red); (3) engineering of genes, genetic elements and genetic variation at increasing scale (orange); and (4) systematic tissue-specific surveys of regulatory elements and transcription (grey). scRNA-seq, single-cell RNA-sequencing analysis; ChIA-PET, chromatin interaction analysis by paired-end tag sequencing; ChIP–seq, chromatin immunoprecipitation followed by sequencing; FAIRE-seq, formaldehyde-assisted isolation of regulatory elements with sequencing; DHS-seq, DNase I-hypersensitive sites sequencing; ATAC-seq, assay for transposase-accessible chromatin using sequencing; MPRA, massively parallel reporter assay; STARR-seq, self-transcribing active regulatory region sequencing; CNN: convolutional neural networks. For further details and primary literature on many of these assays, see ref. [173].

detect de novo variation in dominant disorders and phase biallelic rare variants in recessive disease[13].

The transition from targeted gene tests to genomic sequencing enables recursive reanalysis, including reinterpretation of individual sequences on the basis of subsequent discoveries regarding causal disease alleles and their phenotypic consequences[137]. However, improved molecular diagnostics are required to ensure reliable detection of a subset of genetic disorders, including those arising from triplet repeats and complex rearrangements[138]. Deep sequencing of affected tissues for mosaic variants and the use of RNA sequencing to detect noncoding variants that drive early-onset disease (for example, through effects on splicing) represent new fronts for clinical diagnostics[30].

Other examples of the rapid adoption of new genomic technologies include noninvasive prenatal testing (more than ten million tests by 2018 across multiple countries[139–141]) and the use of recessive carrier panels for couples planning pregnancies. Newborn screening is now universal in many countries, although it is limited to disorders combining high-throughput low-cost detection with effective early interventions (such as diet restrictions or enzyme replacement)[142]. Genetic diagnostics are also increasingly applied to newborn screening as a reflex test following an abnormal (for example, metabolic) screening test[143]. Over the next decade, the repertoire of disorders captured by neonatal screening and prenatal testing is likely to expand markedly. Whereas prenatal testing may be more effective at avoiding disease, the associated ethical issues are more complex[144].

Although genetic testing for rare disease and cancer has exploded, there has been more limited uptake of genetic information in other aspects of healthcare. For example, despite multiple examples of clinically important genetic markers related to drug efficacy and side-effect profile[145], the roll-out of pharmacogenetics has been hampered by a range of factors, including lack of clinical decision support in electronic medical systems to guide the drug choice or dosing by the physician. This has been compounded by challenges in diagnostic testing: complex haplotype structures and structural variants at some key drug metabolism loci necessitate genome sequencing or specific targeted panels to detect all clinically relevant variants.

For common diseases, translational attention is currently focused on the clinical potential of polygenic risk scores. The development of robust polygenic scores for several common diseases has been catalysed by more precise per-variant effect estimates from larger GWAS datasets, improved algorithms for combining information across millions of single-nucleotide polymorphisms, and large-scale biobanks that support score validation[69,146,147]. For example, a genome-wide polygenic score for heart attack, incorporating 6.6 million variants, indicates that 5% of European-descent individuals have a risk of future cardiac events equivalent to that seen in those with less frequent monogenic forms of hypercholesterolaemia[69]. Increasingly, the shift from array-based genotyping to sequence-based analysis is facilitating risk prediction, which integrates information from rare, large-effect alleles with that from polygenic scores[93]. By improving the capture of genetic risk, particularly in non-European populations, and integrating environmental and biomarker data to quantify aspects of non-genetic risk, it should be possible to achieve increasingly accurate prediction of individual disease risk, and to use this information to tailor screening, prevention and treatment. Success will depend on developing models of risk that robustly integrate these diverse data types and on optimizing the strategies deployed to ensure effective implementation.

Table 1 | Examples showing progress in the clinical utility of genetic testing

Mode of clinical utility	Disorder	Gene(s)	Treatment or action
Targeted therapeutics	Cystic fibrosis	*CFTR* (p.G551D)	Ivacaftor
	Retinal dystrophy	*RPE65*	Voretigene neparvovec-rzyl
	Fabry disease	*GLA* (specific variants)	Migalastat
	Spinal muscular atrophy	*SMN1*	Nusinersen
	Familial hypercholesterolaemia	*PCSK9*	Evolocumab and alirocumab
Altering diet or behaviour	Phenylketonuria	*PAH*	Phenylalanine-free diet
	Hereditary fructose intolerance	*ALDOB*	Elimination of fructose and sucrose from diet
	Biotinidase deficiency	*BTD*	Biotin vitamin supplements
	Usher syndrome	*MYO7A* and others	Avoid cycling and other activities that depend on balance and peripheral vision
Prophylactic interventions	Sickle-cell disease	HBB^{ss}	Antibiotics and pneumococcal vaccine
	Brugada syndrome	*SCN5A*	Implantable cardioverter–defibrillator
	Hereditary breast or ovarian cancer	*BRCA1* or *BRCA2*	Bilateral salpingo-oophorectomy
	Multiple endocrine neoplasia 2	*RET*	Thyroidectomy
Increased surveillance	Lynch syndrome	*MSH2* and others	Early initiation of colonoscopy
	Marfan syndrome	*FBN1*	Echocardiography
	Haemochromatosis	*HFE*	Iron screening
	Alpha-1 antitrypsin deficiency	*AAT*	Spirometry, liver function tests

There is an increasing number of examples in which a specific molecular diagnosis results in downstream actions to improve patient outcomes. These actions can range from a change in diet or behaviour to prophylactic surgery. For each example shown, we provide the disorder, the gene and corresponding actions. The GeneReviews compendium (https://www.ncbi.nlm.nih.gov/books/NBK1116/) provides up-to-date information on the management of these and other Mendelian disorders.

The absence of evidence-based guidelines to support healthcare recommendations continues to hinder the clinical applications of genetic data. In some countries, this is compounded by confusion over reimbursement and disparities in testing across society[148]. Many healthcare professionals lack experience in genomic medicine and need education and guidance to practice in the rapidly evolving space of genetic and genomic testing[149]. One consequence of these difficulties has been an expanding direct-to-consumer testing market, variably controlled by country-specific regulations[150], which is moving beyond a focus on ancestry and personal traits, towards models in which individuals have direct access to ordering physicians and genetic counselors[151]. The risk of commercial influence in this model remains high. There are concerns about the consequences of unfettered release of genetic data of dubious or inflated clinical relevance, and limited infrastructure to pull these results into mainstream medical systems.

These advances have fostered debate about the value of genetics for population screening, for both monogenic and complex disorders. Population screening for monogenic disorders is most likely to be initiated for conditions for which risk estimates are well-understood and there are actionable interventions (for example, Lynch syndrome and familial hypercholesterolaemia). Expansion to other disorders requires better understanding of the penetrance of pathogenic alleles in unselected populations[152] and caution before extending screening to longer lists of genes that are less securely implicated in disease causation[153]. As certain countries consider universal capture of genome-wide genetic data at birth or later in life, key questions concern the strategies for releasing this information to citizens and their medical teams to support individual healthcare.

Ultimately, barriers to genomic medicine are most directly overcome by demonstrating clinical utility in disease management and therapeutic decision-making, with evidence for improved patient outcomes. Hereditary cancers provide multiple examples, such as the use of *BRCA1/BRCA2* testing to inform PARP inhibitor treatment in patients with cancer[154]. There is a growing list of diseases for which a molecular diagnosis results in specific interventions designed to improve patient outcomes (https://www.ncbi.nlm.nih.gov/books/NBK1116/) (some examples are listed in Table 1), and there are currently more than 50 FDA-approved drugs for genetic disorders[155]. Although gene therapy has been slow to evolve since its early introduction, recent advances in gene editing are reinvigorating approaches to treat disorders by manipulation of the underlying genetic defects[156].

Looking forward

Over the coming decade, the challenge will be to optimize and to implement at scale, strategies that use human genetics to further the understanding of health and disease, and to maximize the clinical benefit of those discoveries. Realizing these goals will require the concerted effort of researchers in academia and industry to bring about transformational change across a range of highly interconnected domains, for example, through the auspices of the recently established International Common Disease Alliance (https://www.icda.bio). Such efforts will be directed towards establishing: (a) comprehensive inventories of genotype–phenotype relationships across populations and environments; (b) systematic assays of variant- and gene-level function across cell types, states and exposures; (c) improved scalable strategies for turning this basic knowledge into fully developed molecular, cellular and physiological models of disease pathogenesis; and (d) application of those biological insights to drive novel preventative and therapeutic options.

The first of these will involve documenting the full spectrum of natural genetic variation across all human populations, including capture of structural variants, and somatic mutations that accumulate with aging[157,158], and associating these variations with the ever-richer disease-related intermediate and clinical traits available through biobanks and electronic health records. It will be particularly important to include populations historically under-represented in genomic research, following the pioneering work of the H3Africa consortium[159]. As over time, clinically sequenced genomes will outnumber those collected in academia, research and healthcare communities will need to develop a harmonized approach to genomics to transcend historical boundaries. Progress will be critically dependent on platforms and governance that lower barriers to the integration of genetic and phenotypic data across

Box 1

Global genomics

Present and future advances in genetics and genomics have the potential to provide benefits to individuals and societies across the world, but equitable and fair access to those benefits will require proactive measures to address entrenched disparities in scientific capacity and clinical opportunities. This includes:

(1) Global characterization of genetic variation. Systematic catalogues of human genome variation from a broad range of global populations will maximize the genetic diversity available for genetic discovery and clinical implementation. This will support more accurate imputation, enable more effective use of polygenic scores, and improve detection and interpretation of rare large-effect alleles.

(2) Genetic discovery for diseases with restricted geographical coverage. Genetic studies conducted in diverse populations will support genetic discovery as well as functional characterization of hitherto-neglected diseases with disproportionate regional impact and may lead to novel preventative and therapeutic strategies.

(3) An ethical framework for global research. Genetic research needs to proceed on the expectation that the benefits of research will be available to those who have participated, and that relationships between researchers and patients are based on robust expressions of accountability and governance. Participation should be based on appropriate informed individual consent in the context of community involvement, with mechanisms to support equitable data sharing with researchers in participating countries.

(4) Support for local research capacity. Collection of biological specimens and data are dependent on adequate support for data collection, generation and analysis in participating countries, with infrastructure, training and career structures in place to support local researchers and equitable credit for shared research outputs.

(5) Equitable translation. Participants and their communities should have fair and equitable access to the biomedical information arising from the studies in which they are involved to support population-relevant clinical diagnostics and implementation (for example, of polygenic scores). There should be commitment to the development of technologies and strategies that enable the clinical benefits of genomic medicine to be accessible to those in less-developed healthcare environments.

studies and countries, along with technical standards that are reliable, secure and compatible with the international regulatory landscape[160].

Mechanistic interpretation of genetic associations, particularly those in regulatory regions, will be driven by the systematic annotation of sequence variants and genes for functional impact across disease-relevant cell types, enabling mapping of processes contributing to disease development with respect to place (tissue and cell type), time (developmental stage) and context (external influences)[161]. Accelerating efforts to characterize the cellular composition of tissues through single-cell assays[115] will increase the granularity of these observations. Large-scale perturbation studies across diverse cellular and animal models will, together with analyses of coding variants in humans[53,54], provide confidence in causal inference. Large-scale proteomic and metabolomic analyses (in tissues and biological fluids) will provide a bridge to downstream pathways[79,80]. Research access to such functional data, generated at scale, should lower the barriers to mechanistic inference, provide system-wide context and enable researchers to focus wet-laboratory validation on the most critical experiments. Collectively, these efforts will support compilation of a systematic catalogue of key networks and processes that influence normal physiology and disease development and inform a revised molecular taxonomy of disease.

This knowledge will reinforce the essential contribution of human genetics to the identification and prioritization of targets for therapeutic development[89,162]. Insights into the efficacy of target perturbation and potential for adverse events, allied to characterization of translatable biomarkers, provide ways to boost the efficiency of drug-development pipelines[162]. Given the clinical importance of slowing disease progression[163], target-discovery efforts will increasingly need to embrace the genetics of disease progression and treatment response, as these may involve processes distinct from those captured by studies of disease onset.

In parallel, the clinical use of human genetics will benefit from progress towards universal determination of individual genome sequences built through a combination of biobank expansion and direct access within healthcare systems. This will power clinical applications that extend beyond the current focus on neonatal sequencing, Mendelian diagnostics and somatic tumour sequencing[164]. In particular, improvements in polygenic score derivation will boost risk prediction for multifactorial traits, provide a molecular basis for disease classification, support biomarker discovery and therapeutic optimization and contribute to understanding of the variable penetrance of monogenic conditions[69]. Implementing genomic medicine as a routine component of clinical care across diverse healthcare environments will inevitably require investment in the training of healthcare professionals and attention to optimal strategies for returning genetic findings to patients.

The limited heritability of many multifactorial traits constrains the clinical precision available from genetic data alone. This will drive efforts to integrate information on personal environment, lifestyle and behaviour, and to combine prognostic, predictive information on disease risk with longitudinal measures of molecular and clinical state that track an individual's journey from health to disease. Human genetics will also, given its unique potential for causal inference, support identification of the non-genetic risk factors (often modifiable) that directly contribute to disease predisposition and development[165]. As polygenic score performance improves, analysis of individuals who show marked divergence between genetic predisposition and real-world clinical outcomes should define exposures (such as lifestyle choices or gut microbiome) the contribution of which to disease causation remains unclear[166].

Collectively, these developments can be expected to accelerate personalization of healthcare delivery. Provided costs are sustainable, a more preventative perspective on health could emerge, managed through proactive genomic, clinical and lifestyle surveillance using risk scores, complex biomarkers, liquid biopsies and wearables. Improved understanding of aetiological heterogeneity, patterns of sharing of genetic risk across diseases, variation in therapeutic response and risk of adverse events will enhance targeting of preventative and therapeutic interventions[167]. At the population level, intervention strategies will seek to combine population-wide and targeted strategies to best effect[168]. It will be critical to ensure that these benefits are available to as many as possible, so that genomics reduces, rather than exacerbates, national and global health disparities[55,169] (Box 1).

The developments described above, represent variations on the theme of 'reading' the genome. The emerging capacity to block this reading (for example, through siRNA therapies[170]) or even to 'write' the genome (through CRISPR editing) promises to be equally transformative, providing new opportunities to correct, and even cure, Mendelian disease. Spectacular advances in developing novel therapeutic strategies are likely for many diseases, based, for example, on ex vivo cellular manipulation[171] or in vivo somatic cell editing[172].

Importantly, developments in genomic medicine need to proceed in a bioethical framework for research and clinical use that recognizes

the personal relevance of human genetics and the critical importance of autonomous consent and the protection of privacy, while minimizing the adverse consequences of genetic exceptionalism. Governance needs to reaffirm the rights of citizens to make individual contributions to scientific progress through research participation and encourage the responsible exchange of data for clinical and research purposes.

Future prospects

Over the past two decades, understanding of the genetic basis of human disease has been transformed by a combination of spectacular technological and analytical advances, collaborative commitment to the development of foundational resources and the collection and analysis of vast amounts of genetic, molecular and clinical data. The biological insights derived from these data are, increasingly, drivers of translational innovation, and widening personal access to large-scale genetic and molecular data promises to reshape medical care.

However, for the full potential of genomic medicine to be realized, there will need to be sustained collaborative endeavour on several fronts to ensure that the capacity to generate ever more detailed maps of the relationships between sequence variation and biomedical phenotypes delivers a comprehensive understanding of disease mechanisms that can be translated into the medicines of tomorrow.

1. International Human Genome Sequencing Consortium. Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921 (2001).
This paper describes the first analyses from the draft human genome sequence assembled over the previous decade: it launched modern human genetics and represents a tribute to the power of collaborative science.
2. International HapMap Consortium. The International HapMap Project. *Nature* **426**, 789–796 (2003).
The HapMap Consortium developed the first genome-wide maps of common sequence variation, using this information to lay out the haplotypic structure of this variation across three major ancestral groupings (from Europe, East Asia and Africa).
3. 1000 Genomes Project Consortium. A global reference for human genetic variation. *Nature* **526**, 68–74 (2015).
4. The Haplotype Reference Consortium. A reference panel of 64,976 haplotypes for genotype imputation. *Nat. Genet.* **48**, 1279–1283 (2016).
5. Lek, M. et al. Analysis of protein-coding genetic variation in 60,706 humans. *Nature* **536**, 285–291 (2016).
6. Karczewski, K. J. et al. Variation across 141,456 human exomes and genomes reveals the spectrum of loss-of-function intolerance across human protein-coding genes. Preprint at *bioRxiv* https://doi.org/10.1101/531210 (2019).
The most recent report from the genome aggregation database (gnomAD) project combining human sequence data on an unprecedented scale to characterize rare, high-impact variants and their relationship to health and disease.
7. Kremer, B. et al. A worldwide study of the Huntington's disease mutation. The sensitivity and specificity of measuring CAG repeats. *N. Engl. J. Med.* **330**, 1401–1406 (1994).
8. Collins, F. S. Identifying human disease genes by positional cloning. *Harvey Lect.* **86**, 149–164 (1990–1991).
9. Gusella, J. F. & MacDonald, M. E. Huntington's disease and repeating trinucleotides. *N. Engl. J. Med.* **330**, 1450–1451 (1994).
10. Vissers, L. E. et al. Array-based comparative genomic hybridization for the genomewide detection of submicroscopic chromosomal abnormalities. *Am. J. Hum. Genet.* **73**, 1261–1270 (2003).
11. Ng, S. B. et al. Targeted capture and massively parallel sequencing of 12 human exomes. *Nature* **461**, 272–276 (2009).
12. Wheeler, D. A. et al. The complete genome of an individual by massively parallel DNA sequencing. *Nature* **452**, 872–876 (2008).
13. Vissers, L. E. et al. A de novo paradigm for mental retardation. *Nat. Genet.* **42**, 1109–1112 (2010).
14. Deciphering Developmental Disorders Study. Prevalence and architecture of de novo mutations in developmental disorders. *Nature* **542**, 433–438 (2017).
15. Splinter, K. et al. Effect of genetic diagnosis on patients with previously undiagnosed disease. *N. Engl. J. Med.* **379**, 2131–2139 (2018).
16. Deciphering Developmental Disorders Study. Large-scale discovery of novel genetic causes of developmental disorders. *Nature* **519**, 223–228 (2015).
17. Firth, H. V. et al. DECIPHER: database of chromosomal imbalance and phenotype in humans using ensembl resources. *Am. J. Hum. Genet.* **84**, 524–533 (2009).
18. Philippakis, A.A. et al. The Matchmaker Exchange: a platform for rare disease gene discovery. *Hum. Mutat.* **36**, 915–921 (2015).
19. Boycott, K. M. et al. International cooperation to enable the diagnosis of all rare genetic diseases. *Am. J. Hum. Genet.* **100**, 695–705 (2017).
20. Kennedy, J. et al. KAT6A syndrome: genotype–phenotype correlation in 76 patients with pathogenic *KAT6A* variants. *Genet. Med.* **21**, 850–860 (2019).
21. Amberger, J. S., Bocchini, C. A., Schiettecatte, F., Scott, A. F. & Hamosh, A. OMIM.org: Online Mendelian Inheritance in Man (OMIM), an online catalog of human genes and genetic disorders. *Nucleic Acids Res.* **43**, D789–D798 (2015).
22. Landrum, M. J. et al. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res.* **42**, D980–D985 (2014).
23. Rehm, H. L. et al. ClinGen—the Clinical Genome Resource. *N. Engl. J. Med.* **372**, 2235–2242 (2015).
24. Minikel, E. V. et al. Quantifying prion disease penetrance using large population control cohorts. *Sci. Transl. Med.* **8**, 322ra9 (2016).
25. Flannick, J. et al. Assessing the phenotypic effects in the general population of rare variants in genes for a dominant Mendelian form of diabetes. *Nat. Genet.* **45**, 1380–1385 (2013).
26. Bečanović, K. et al. A SNP in the *HTT* promoter alters NF-κB binding and is a bidirectional genetic modifier of Huntington disease. *Nat. Neurosci.* **18**, 807–816 (2015).
27. Castel, S. E. et al. Modified penetrance of coding variants by *cis*-regulatory variation contributes to disease risk. *Nat. Genet.* **50**, 1327–1334 (2018).
28. MacArthur, D. G. et al. Guidelines for investigating causality of sequence variants in human disease. *Nature* **508**, 469–476 (2014).
29. Findlay, G. M. et al. Accurate classification of *BRCA1* variants with saturation genome editing. *Nature* **562**, 217–222 (2018).
This paper describes an effort to obtain massively parallel functional assessments of all potential sequence variants in a gene causal for familial breast cancer, and thereby provide more confident predictions of clinical significance for those found to carry a previously unseen mutation.
30. Cummings, B. B. et al. Improving genetic diagnosis in Mendelian disease with transcriptome sequencing. *Sci. Transl. Med.* **9**, eaal5209 (2017).
31. Butcher, D. T. et al. CHARGE and Kabuki syndromes: gene-specific DNA methylation signatures identify epigenetic mechanisms linking these clinically overlapping conditions. *Am. J. Hum. Genet.* **100**, 773–788 (2017).
32. Short, P. J. et al. De novo mutations in regulatory elements in neurodevelopmental disorders. *Nature* **555**, 611–616 (2018).
33. Jaganathan, K. et al. Predicting splicing from primary sequence with deep learning. *Cell* **176**, 535–548 (2019).
34. Hugot, J. P. et al. Association of *NOD2* leucine-rich repeat variants with susceptibility to Crohn's disease. *Nature* **411**, 599–603 (2001).
35. Ogura, Y. et al. A frameshift mutation in *NOD2* associated with susceptibility to Crohn's disease. *Nature* **411**, 603–606 (2001).
36. Barbosa, J., Chern, M. M., Noreen, H. & Anderson, V. E. Analysis of linkage between the major histocompatibility system and juvenile, insulin-dependent diabetes in multiplex families. Reanalysis of data. *J. Clin. Invest.* **62**, 492–495 (1978).
37. Risch, N. & Merikangas, K. The future of genetic studies of complex human diseases. *Science* **273**, 1516–1517 (1996).
A highly influential commentary that switched much of the complex trait genetics field from linkage to association and imagined GWAS several years before the technology made them a reality.
38. Klein, R. J. et al. Complement factor H polymorphism in age-related macular degeneration. *Science* **308**, 385–389 (2005).
Arguably the first full GWAS, this study demonstrated the potential of agnostic genomic surveys to highlight entirely novel biology—in this case, a role for complement in the pathogenesis of macular degeneration.
39. Duerr, R. H. et al. A genome-wide association study identifies *IL23R* as an inflammatory bowel disease gene. *Science* **314**, 1461–1463 (2006).
40. Wellcome Trust Case Control Consortium. Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. *Nature* **447**, 661–678 (2007).
This paper, which describes what was, at the time, the largest GWAS yet conducted, demonstrated the broad applicability of the approach, and set the scene for the decade of GWAS discovery that followed.
41. Mahajan, A. et al. Fine-mapping type 2 diabetes loci to single-variant resolution using high-density imputation and islet-specific epigenome maps. *Nat. Genet.* **50**, 1505–1513 (2018).
42. Astle, W. J. et al. The allelic landscape of human blood cell trait variation and links to common complex disease. *Cell* **167**, 1415–1429 (2016).
43. Lango Allen, H. et al. Hundreds of variants clustered in genomic loci and biological pathways affect human height. *Nature* **467**, 832–838 (2010).
44. Buniello, A. et al. The NHGRI–EBI GWAS catalog of published genome-wide association studies, targeted arrays and summary statistics 2019. *Nucleic Acids Res.* **47**, D1005–D1012 (2019).
45. Fuchsberger, C. et al. The genetic architecture of type 2 diabetes. *Nature* **536**, 41–47 (2016).
46. UK10K Consortium. The UK10K project identifies rare variants in health and disease. *Nature* **526**, 82–90 (2015).
47. Manolio, T. A. et al. Finding the missing heritability of complex diseases. *Nature* **461**, 747–753 (2009).
48. Wainschtein, P. et al. Recovery of trait heritability from whole genome sequence data. Preprint at *bioRxiv* https://doi.org/10.1101/588020 (2019).
49. Flannick, J. et al. Loss-of-function mutations in *SLC30A8* protect against type 2 diabetes. *Nat. Genet.* **46**, 357–363 (2014).
50. Moltke, I. et al. A common Greenlandic *TBC1D4* variant confers muscle insulin resistance and type 2 diabetes. *Nature* **512**, 190–193 (2014).
51. Nioi, P. et al. Variant *ASGR1* associated with a reduced risk of coronary artery disease. *N. Engl. J. Med.* **374**, 2131–2141 (2016).
52. Pollin, T. I. et al. A null mutation in human *APOC3* confers a favorable plasma lipid profile and apparent cardioprotection. *Science* **322**, 1702–1705 (2008).
53. Narasimhan, V. M. et al. Health and population effects of rare gene knockouts in adult humans with related parents. *Science* **352**, 474–477 (2016).
A powerful illustration of the insights that can be gained from the study of individuals who are homozygous carriers of null alleles in genes of biomedical interest.
54. Saleheen, D. et al. Human knockouts and phenotypic analysis in a cohort with a high rate of consanguinity. *Nature* **544**, 235–239 (2017).

55. Martin, A. R. et al. Clinical use of current polygenic risk scores may exacerbate health disparities. *Nat. Genet*. **51**, 584–591 (2019).
An elegant distillation of the ways in which the historical focus of genetic discovery among populations of European descent has adverse consequences for both discovery and translation.
56. Zeng, J. et al. Signatures of negative selection in the genetic architecture of human complex traits. *Nat. Genet*. **50**, 746–753 (2018).
57. Genovese, G. et al. Association of trypanolytic ApoL1 variants with kidney disease in African Americans. *Science* **329**, 841–845 (2010).
58. Luzzatto, L., Usanga, F. A. & Reddy, S. Glucose-6-phosphate dehydrogenase deficient red cells: resistance to infection by malarial parasites. *Science* **164**, 839–842 (1969).
59. Wojcik, G. L. et al. Genetic analyses of diverse populations improves discovery for complex traits. *Nature* **570**, 514–518 (2019).
60. Farh, K. K. et al. Genetic and epigenetic fine mapping of causal autoimmune disease variants. *Nature* **518**, 337–343 (2015).
An early demonstration of the value of overlapping genome-wide association data with tissue-specific regulatory maps to define the cell-types and tissues likely to be driving disease pathology.
61. Freund, M. K. et al. Phenotype-specific enrichment of Mendelian disorder genes near GWAS regions across 62 complex traits. *Am. J. Hum. Genet*. **103**, 535–552 (2018).
62. Boyle, E. A., Li, Y. I. & Pritchard, J. K. An expanded view of complex traits: from polygenic to omnigenic. *Cell* **169**, 1177–1186 (2017).
63. de Lange, K. M. et al. Genome-wide association study implicates immune activation of multiple integrin genes in inflammatory bowel disease. *Nat. Genet*. **49**, 256–261 (2017).
64. Sekar, A. et al. Schizophrenia risk from complex variation of complement component 4. *Nature* **530**, 177–183 (2016).
65. Jostins, L. et al. Host–microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
66. Richardson, T. G., Harrison, S., Hemani, G. & Davey Smith, G. An atlas of polygenic risk score associations to highlight putative causal relationships across the human phenome. *eLife* **8**, e43657 (2019).
67. Dendrou, C. A. et al. Resolving TYK2 locus genotype-to-phenotype differences in autoimmunity. *Sci. Transl. Med*. **8**, 363ra149 (2016).
68. Voight, B. F. et al. Plasma HDL cholesterol and risk of myocardial infarction: a Mendelian randomisation study. *Lancet* **380**, 572–580 (2012).
69. Khera, A. V. et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat. Genet*. **50**, 1219–1224 (2018).
This paper used large-scale genetic data to demonstrate the clinical potential of the polygenic scores that can be constructed for many common diseases, emphasizing that, in some situations, the lifetime risk of disease for those with the highest scores approaches that of established monogenic disease.
70. Mahmood, S. S., Levy, D., Vasan, R. S. & Wang, T. J. The Framingham Heart Study and the epidemiology of cardiovascular disease: a historical perspective. *Lancet* **383**, 999–1008 (2014).
71. Gudbjartsson, D. F. et al. Variants conferring risk of atrial fibrillation on chromosome 4q25. *Nature* **448**, 353–357 (2007).
72. Thorgeirsson, T. E. et al. A variant associated with nicotine dependence, lung cancer and peripheral arterial disease. *Nature* **452**, 638–642 (2008).
73. Collins, F. S. & Varmus, H. A new initiative on precision medicine. *N. Engl. J. Med*. **372**, 793–795 (2015).
74. Bycroft, C. et al. The UK Biobank resource with deep phenotyping and genomic data. *Nature* **562**, 203–209 (2018).
75. Nagai, A. et al. Overview of the BioBank Japan Project: study design and profile. *J. Epidemiol*. **27**, S2–S8 (2017).
76. Chen, Z. et al. China Kadoorie Biobank of 0.5 million people: survey methods, baseline characteristics and long-term follow-up. *Int. J. Epidemiol*. **40**, 1652–1666 (2011).
77. Kohane, I. S. Using electronic health records to drive discovery in disease genomics. *Nat. Rev. Genet*. **12**, 417–428 (2011).
78. Abul-Husn, N. S. & Kenny, E. E. Personalized medicine and the power of electronic health records. *Cell* **177**, 58–69 (2019).
79. Sun, B. B. et al. Genomic atlas of the human plasma proteome. *Nature* **558**, 73–79 (2018).
80. Shin, S. Y. et al. An atlas of genetic influences on human blood metabolites. *Nat. Genet*. **46**, 543–550 (2014).
81. Sung, Y. J. et al. A multi-ancestry genome-wide study incorporating gene-smoking interactions identifies multiple new loci for pulse pressure and mean arterial pressure. *Hum. Mol. Genet*. **28**, ddz070 (2019).
82. Pendergrass, S. A. & Crawford, D. C. Using electronic health records to generate phenotypes for research. *Curr. Protoc. Hum. Genet*. **100**, e80 (2019).
83. Denny, J. C. et al. PheWAS: demonstrating the feasibility of a phenome-wide scan to discover gene–disease associations. *Bioinformatics* **26**, 1205–1210 (2010).
84. Liu, Z. et al. Mendelian randomization analysis dissects the relationship between NAFLD, T2D and obesity and provides implications for precision medicine. Preprint at *bioRxiv* https://doi.org/10.1101/657734 (2019).
85. Cotsapas, C. et al. Pervasive sharing of genetic effects in autoimmune disease. *PLoS Genet*. **7**, e1002254 (2011).
86. Kanai, M. et al. Genetic analysis of quantitative traits in the Japanese population links cell types to complex human diseases. *Nat. Genet*. **50**, 390–400 (2018).
87. Gudmundsson, J. et al. Two variants on chromosome 17 confer prostate cancer risk, and the one in *TCF2* protects against type 2 diabetes. *Nat. Genet*. **39**, 977–983 (2007).
88. Abul-Husn, N. S. et al. A protein-truncating *HSD17B13* variant and protection from chronic liver disease. *N. Engl. J. Med*. **378**, 1096–1106 (2018).
89. Plenge, R. M., Scolnick, E. M. & Altshuler, D. Validating therapeutic targets through human genetics. *Nat. Rev. Drug Discov*. **12**, 581–594 (2013).
90. Nielsen, J. B. et al. Biobank-driven genomic discovery yields new insight into atrial fibrillation biology. *Nat. Genet*. **50**, 1234–1239 (2018).
91. Flannick, J. et al. Exome sequencing of 20,791 cases of type 2 diabetes and 24,440 controls. *Nature* **570**, 71–76 (2019).
92. Van Hout, C. V. et al. Whole exome sequencing and characterization of coding variation in 49,960 individuals in the UK Biobank. Preprint at *bioRxiv* https://doi.org/10.1101/572347 (2019).
93. Khera, A. V. et al. Whole-genome sequencing to characterize monogenic and polygenic contributions in patients hospitalized with early-onset myocardial infarction. *Circulation* **139**, 1593–1602 (2019).
94. Grarup, N. et al. Loss-of-function variants in *ADCY3* increase risk of obesity and type 2 diabetes. *Nat. Genet*. **50**, 172–174 (2018).
95. Cohen, J. C., Boerwinkle, E., Mosley, T. H. Jr & Hobbs, H. H. Sequence variations in *PCSK9*, low LDL, and protection against coronary heart disease. *N. Engl. J. Med*. **354**, 1264–1272 (2006).
96. Rivas, M. A. et al. Deep resequencing of GWAS loci identifies independent rare variants associated with inflammatory bowel disease. *Nat. Genet*. **43**, 1066–1073 (2011).
97. Chong, J. X. et al. The genetic basis of Mendelian phenotypes: discoveries, challenges, and opportunities. *Am. J. Hum. Genet*. **97**, 199–215 (2015).
98. Goldstein, J. L. & Brown, M. S. A century of cholesterol and coronaries: from plaques to genes to statins. *Cell* **161**, 161–172 (2015).
99. Peltonen, L., Perola, M., Naukkarinen, J. & Palotie, A. Lessons from studying monogenic disease for common disease. *Hum. Mol. Genet*. **15**, R67–R74 (2006).
100. Bulik-Sullivan, B. et al. An atlas of genetic correlations across human diseases and traits. *Nat. Genet*. **47**, 1236–1241 (2015).
101. Musunuru, K. et al. From noncoding variant to phenotype via *SORT1* at the 1p13 cholesterol locus. *Nature* **466**, 714–719 (2010).
Widely accepted as the first paper to describe the detailed mechanistic dissection of a complex trait locus and still one of relatively few examples of success in this endeavour.
102. Claussnitzer, M. et al. FTO obesity variant circuitry and adipocyte browning in humans. *N. Engl. J. Med*. **373**, 895–907 (2015).
103. Smemo, S. et al. Obesity-associated variants within *FTO* form long-range functional connections with *IRX3*. *Nature* **507**, 371–375 (2014).
104. Gupta, R. M. et al. A genetic variant associated with five vascular diseases is a distal regulator of endothelin-1 gene expression. *Cell* **170**, 522–533 (2017).
105. Small, K. S. et al. Regulatory variants at *KLF14* influence type 2 diabetes risk via a female-specific effect on adipocyte size and body composition. *Nat. Genet*. **50**, 572–580 (2018).
106. Lin, Y. T. et al. *APOE4* causes widespread molecular and cellular alterations associated with Alzheimer's disease phenotypes in human iPSC-derived brain cell types. *Neuron* **98**, 1294 (2018).
107. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
108. Roadmap Epigenomics Consortium et al. Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
The flagship paper from the NIH Roadmap Epigenomics Consortium, describing a detailed analysis of the epigenomic profiles of more than 100 human cell types and characterizing how these regulatory patterns relate to gene regulation, cellular differentiation and human disease.
109. Tansey, K. E., Cameron, D. & Hill, M. J. Genetic risk for Alzheimer's disease is concentrated in specific macrophage and microglial transcriptional networks. *Genome Med*. **10**, 14 (2018).
110. Gamazon, E. R. et al. Using an atlas of gene regulation across 44 human tissues to inform complex disease- and trait-associated variation. *Nat. Genet*. **50**, 956–967 (2018).
111. Ongen, H. et al. Estimating the causal tissues for complex traits and diseases. *Nat. Genet*. **49**, 1676–1683 (2017).
112. GTEx Consortium. Genetic effects on gene expression across human tissues. *Nature* **550**, 204–213 (2017).
One of several papers describing the results obtained from the Genotype-Tissue Expression (GTEx) project, which used RNA-sequence data from more than 40 human tissues in several hundred individuals to explore the relationship between DNA sequence variation and tissue-specific expression.
113. Chiou, J. et al. Single cell chromatin accessibility reveals pancreatic islet cell type- and state-specific regulatory programs of diabetes risk. Preprint at *bioRxiv* https://doi.org/10.1101/693671 (2019).
114. HuBMAP Consortium. The human body at cellular resolution: the NIH Human Biomolecular Atlas Program. *Nature* **574**, 187–192 (2019).
115. Regev, A. et al. The Human Cell Atlas. *eLife* **6**, e27041 (2017).
116. Lloyd, K. C. A knockout mouse resource for the biomedical research community. *Ann. NY Acad. Sci*. **1245**, 24–26 (2011).
117. Finucane, H. K. et al. Heritability enrichment of specifically expressed genes identifies disease-relevant tissues and cell types. *Nat. Genet*. **50**, 621–629 (2018).
118. Delaneau, O. et al. Chromatin three-dimensional interactions mediate genetic effects on gene expression. *Science* **364**, eaat8266 (2019).
119. Zhou, J. & Troyanskaya, O. G. Predicting effects of noncoding variants with deep learning-based sequence model. *Nat. Methods* **12**, 931–934 (2015).
120. Kelley, D. R., Snoek, J. & Rinn, J. L. Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Res*. **26**, 990–999 (2016).
121. Patwardhan, R. P. et al. High-resolution analysis of DNA regulatory elements by synthetic saturation mutagenesis. *Nat. Biotechnol*. **27**, 1173–1175 (2009).
122. Melnikov, A. et al. Systematic dissection and optimization of inducible enhancers in human cells using a massively parallel reporter assay. *Nat. Biotechnol*. **30**, 271–277 (2012).
123. Wang, X. et al. High-resolution genome-wide functional dissection of transcriptional regulatory regions and nucleotides in human. *Nat. Commun*. **9**, 5380 (2018).
124. Potting, C. et al. Genome-wide CRISPR screen for PARKIN regulators reveals transcriptional repression as a determinant of mitophagy. *Proc. Natl Acad. Sci. USA* **115**, E180–E189 (2018).
125. Gasperini, M. et al. A genome-wide framework for mapping gene regulation via cellular genetic screens. *Cell* **176**, 377–390 (2019).

126. Fulco, C. P. et al. Activity-by-contact model of enhancer-promoter regulation from thousands of CRISPR perturbations. *Nat. Genet*. **51**, 1664–1669 (2019).
127. Komor, A. C., Kim, Y. B., Packer, M. S., Zuris, J. A. & Liu, D. R. Programmable editing of a target base in genomic DNA without double-stranded DNA cleavage. *Nature* **533**, 420–424 (2016).
128. Feldman, D. et al. Optical pooled screens in human cells. *Cell* **179**, 787–799 (2019).
129. Dixit, A. et al. Perturb-Seq: dissecting molecular circuits with scalable single-cell RNA profiling of pooled genetic screens. *Cell* **167**, 1853–1866 (2016).
130. Datlinger, P. et al. Pooled CRISPR screening with single-cell transcriptome readout. *Nat. Methods* **14**, 297–301 (2017).
131. Ding, L. et al. Perspective on oncogenic processes at the end of the beginning of cancer genomics. *Cell* **173**, 305–320 (2018).
132. Chae, Y. K. et al. Path toward precision oncology: review of targeted therapy studies and tools to aid in defining "actionability" of a molecular lesion and patient management support. *Mol. Cancer Ther*. **16**, 2645–2655 (2017).
133. Shearer, A. E. & Smith, R. J. Massively parallel sequencing for genetic diagnosis of hearing loss: the new standard of care. *Otolaryngol. Head Neck Surg*. **153**, 175–182 (2015).
134. Yang, Y. et al. Molecular findings among patients referred for clinical whole-exome sequencing. *J. Am. Med. Assoc*. **312**, 1870–1879 (2014).
135. Friedman, J. M. et al. Genome-wide sequencing in acutely ill infants: genomic medicine's critical application? *Genet. Med*. **21**, 498–504 (2019).
136. Clark, M. M. et al. Diagnosis of genetic diseases in seriously ill children by rapid whole-genome sequencing and automated phenotyping and interpretation. *Sci. Transl. Med*. **11**, eaat6177 (2019).
137. Eldomery, M. K. et al. Lessons learned from additional research analyses of unsolved clinical exome cases. *Genome Med*. **9**, 26 (2017).
138. Boycott, K. M. et al. A diagnosis for all rare genetic diseases: the horizon and the next frontiers. *Cell* **177**, 32–37 (2019).
139. Liu, S. et al. Genomic analyses from non-invasive prenatal testing reveal genetic associations, patterns of viral infections, and Chinese population history. *Cell* **175**, 347–359 (2018).
140. Thung, D. T., Beulen, L., Hehir-Kwa, J. & Faas, B. H. Implementation of whole genome massively parallel sequencing for noninvasive prenatal testing in laboratories. *Expert Rev. Mol. Diagn*. **15**, 111–124 (2015).
141. Lo, J. O., Cori D, F., Norton, M. E. & Caughey, A. B. Noninvasive prenatal testing. *Obstet. Gynecol. Surv*. **69**, 89–99 (2014).
142. Watson, M. S. et al. Newborn screening: toward a uniform screening panel and system—executive summary. *Pediatrics* **117**, S296–S307 (2006).
143. Currier, R. J. et al. Genomic sequencing in cystic fibrosis newborn screening: what works best, two-tier predefined CFTR mutation panels or second-tier CFTR panel followed by third-tier sequencing? *Genet. Med*. **19**, 1159–1163 (2017).
144. Bauer, P. E. "Tell them it's not so bad": prenatal screening for Down syndrome and the bias toward abortion. *Intellect. Dev. Disabil*. **46**, 247–251 (2008).
145. Volpi, S. et al. Research directions in the clinical implementation of pharmacogenomics: an overview of US Programs and Projects. *Clin. Pharmacol. Ther*. **103**, 778–786 (2018).
146. International Schizophrenia Consortium. Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature* **460**, 748–752 (2009).
147. Evans, D. M., Visscher, P. M. & Wray, N. R. Harnessing the information contained within genome-wide association studies to improve individual prediction of complex disease risk. *Hum. Mol. Genet*. **18**, 3525–3531 (2009).
148. Kurian, A. W. et al. Genetic testing and results in a population-based cohort of breast cancer patients and ovarian cancer patients. *J. Clin. Oncol*. **37**, 1305–1315 (2019).
149. Katz, S. J. et al. Association of attending surgeon with variation in the receipt of genetic testing after diagnosis of breast cancer. *JAMA Surg*. **153**, 909–916 (2018).
150. Kalokairinou, L. et al. Legislation of direct-to-consumer genetic testing in Europe: a fragmented regulatory landscape. *J. Community Genet*. **9**, 117–132 (2018).
151. Allyse, M. A., Robinson, D. H., Ferber, M. J. & Sharp, R. R. Direct-to-Consumer Testing 2.0: emerging models of direct-to-consumer genetic testing. *Mayo Clin. Proc*. **93**, 113–120 (2018).
152. Wright, C. F. et al. Assessing the pathogenicity, penetrance, and expressivity of putative disease-causing variants in a population setting. *Am. J. Hum. Genet*. **104**, 275–286 (2019).
153. ACMG Board of Directors. The use of ACMG secondary findings recommendations for general population screening: a policy statement of the American College of Medical Genetics and Genomics (ACMG). *Genet. Med*. **21**, 1467–1468 (2019).
154. George, A., Kaye, S. & Banerjee, S. Delivering widespread BRCA testing and PARP inhibition to patients with ovarian cancer. *Nat. Rev. Clin. Oncol*. **14**, 284–296 (2017).
155. CenterWatch. *FDA Approved Drugs for Genetic Disease* (CenterWatch); https://www.centerwatch.com/drug-information/fda-approved-drugs/therapeutic-area/34/genetic-disease
156. Wu, Y. et al. Highly efficient therapeutic gene editing of human hematopoietic stem cells. *Nat. Med*. **25**, 776–783 (2019).
157. Jaiswal, S. et al. Clonal hematopoiesis and risk of atherosclerotic cardiovascular disease. *N. Engl. J. Med*. **377**, 111–121 (2017).
158. Lee, M. H. et al. Somatic APP gene recombination in Alzheimer's disease and normal neurons. *Nature* **563**, 639–645 (2018).
159. H3Africa Consortium. Enabling the genomic revolution in Africa. *Science* **344**, 1346–1348 (2014).
160. Dolman, L. et al. ClinGen advancing genomic data-sharing standards as a GA4GH driver project. *Hum. Mutat*. **39**, 1686–1689 (2018).
161. Fairfax, B. P. et al. Innate immune activity conditions the effect of regulatory variants upon monocyte gene expression. *Science* **343**, 1246949 (2014).
162. Nelson, M. R. et al. The support of human genetic evidence for approved drug indications. *Nat. Genet*. **47**, 856–860 (2015).
An influential paper that demonstrates that the probability of therapeutic success is enhanced (by around a factor of 2) for drug targets with supportive evidence from human genetics.
163. Veitch, D. P. et al. Understanding disease progression and improving Alzheimer's disease clinical trials: recent highlights from the Alzheimer's Disease Neuroimaging Initiative. *Alzheimers Dement*. **15**, 106–152 (2019).
164. Ashley, E. A. Towards precision medicine. *Nat. Rev. Genet*. **17**, 507–522 (2016).
165. Franks, P. W. & McCarthy, M. I. Exposing the exposures responsible for type 2 diabetes and obesity. *Science* **354**, 69–73 (2016).
166. Sanna, S. et al. Causal relationships among the gut microbiome, short-chain fatty acids and metabolic diseases. *Nat. Genet*. **51**, 600–605 (2019).
167. Udler, M. S. et al. Type 2 diabetes genetic loci informed by multi-trait associations point to disease mechanisms and subtypes: a soft clustering analysis. *PLoS Med*. **15**, e1002654 (2018).
168. Rose, G. Sick individuals and sick populations. *Int. J. Epidemiol*. **14**, 32–38 (1985).
169. Hindorff, L. A. et al. Prioritizing diversity in human genomics research. *Nat. Rev. Genet*. **19**, 175–185 (2018).
170. Wittrup, A. & Lieberman, J. Knocking down disease: a progress report on siRNA therapeutics. *Nat. Rev. Genet*. **16**, 543–552 (2015).
171. Schuster, S. J. et al. Chimeric antigen receptor T cells in refractory B-cell lymphomas. *N. Engl. J. Med*. **377**, 2545–2554 (2017).
172. Ding, Q. et al. Permanent alteration of PCSK9 with in vivo CRISPR–Cas9 genome editing. *Circ. Res*. **115**, 488–492 (2014).
173. Risca, V. I. & Greenleaf, W. J. Unraveling the 3D genome: genomics tools for multiscale exploration. *Trends Genet*. **31**, 357–372 (2015).

Acknowledgements We acknowledge grant funding from the following funders. J.H.C.: NIH (U01DK062429, U01DK062422 and R01DK106593); N.J.C.: NIH (U01HG009086, U54MD010722 and R01MH113362); E.T.D.: Swiss National Science Foundation, Louis Jeantet Foundation; E.E.K.: NIH (R01HG010297, R01HL104608, U01HG009610, R01DK110113, U01HG009080, X01HL1345 and UM1HG0089001); C.M.L.: NIHR Oxford Biomedical Research Centre and NIH (5P50HD028138-27); K.N.N.: NHMRC (APP1113531); S.E.P.: NIH (U01HG006485 and 1U41HG009649); C.N.R.: NIH Intramural Program at the Center for Research on Genomics and Global Health and National Human Genome Research Institute; and M.I.M.: Wellcome (090532, 098381, 106130, 203141 and 212259) and NIH (U01-DK105535, DK085545 and DK098032). Personal funding comes from the following sources. M.C.: Next Generation Fund at the Broad Institute of MIT and Harvard; J.H.C.: Sanford Grossman Charitable Trust and Helmsley Charitable Trust; R.C.: UKBiobank; C.M.L.: Li Ka Shing Foundation; J.S.: Howard Hughes Medical Institute; and M.I.M.: Wellcome and NIHR.

Author contributions M.C. and M.I.M. coordinated the drafting of the Review. All other authors contributed to drafting of the text and approved the final version.

Competing interests R.C. has received research grants from British Heart Foundation, Cancer Research UK, Medical Research Council, Merck & Co, UKBiobank, Wellcome and Medco; a Pfizer Prize Award (to NDPH) and is named on a patent for a statin-related myopathy genetic test. R.C. receives no personal remuneration from these: any share of royalties or other payments have been waived in favour of NDPH. E.T.D. is chairman and board member of Hybridstat and on the advisory board of DNAnexus. M.E.H. is a co-founder, shareholder and director of Congenica. S.K. has received research grants from Bayer and Novartis; is on Scientific Advisory Boards of Regeneron Genetics Center, Corvidia Therapeutics and Maze Therapeutics; has equity in San Therapeutics, Catabasis, Verve and Maze Therapeutics and is a consultant for Maze Therapeutics, Alynlam, ExpertConnect, Leerink Partners, Noble Insights, Bayer and Novo Ventures. E.E.K. receives honoraria from Illumina and Regeneron Pharmaceuticals. C.M.L. has research collaborations with Novo Nordisk and Bayer, receiving no personal payment. D.G.M. is co-founder and shareholder of Goldfinch Bio and has received research funding from AbbVie, Biogen, BioMarin, Merck, Pfizer and Sanofi-Genzyme. S.E.P. is a member of the Scientific Advisory Board of Baylor Genetics. J.S. is a member of the Scientific Advisory Board for Maze Therapeutics, Camp4 Therapeutics, Nanostring, Phase Genomics, Adaptive Biotechnology and Stratos Genomics, a founder of Phase Genomics and a consultant for Guardant Health. M.I.M. was a member of advisory panels for Pfizer, NovoNordisk and Zoe Global; received honoraria from Merck, Pfizer, NovoNordisk and Eli Lilly and research funding from Abbvie, Astra Zeneca, Boehringer Ingelheim, Eli Lilly, Janssen, Merck, NovoNordisk, Pfizer, Roche, Sanofi Aventis, Servier and Takeda. As of June 2019, M.I.M. is an employee of Genentech and a holder of Roche stock. The views expressed in this article are those of the author(s) and not necessarily those of the NHS, the NIHR, or the Department of Health.

Additional information

Correspondence and requests for materials should be addressed to M.I.M.

Peer review information *Nature* thanks Ewan Birney, Jeff Schloss, Richard Trembath and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.
Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© Springer Nature Limited 2020

A repeating fast radio burst source localized to a nearby spiral galaxy

https://doi.org/10.1038/s41586-019-1866-z

Received: 17 October 2019

Accepted: 8 November 2019

Published online: 6 January 2020

B. Marcote[1], K. Nimmo[2,3], J. W. T. Hessels[2,3]*, S. P. Tendulkar[4,5], C. G. Bassa[2], Z. Paragi[1], A. Keimpema[1], M. Bhardwaj[4,5], R. Karuppusamy[6], V. M. Kaspi[4,5], C. J. Law[7], D. Michilli[4,5], K. Aggarwal[8,9], B. Andersen[4,5], A. M. Archibald[3,10], K. Bandura[9,11], G. C. Bower[12], P. J. Boyle[4,5], C. Brar[4,5], S. Burke-Spolaor[8,9], B. J. Butler[13], T. Cassanelli[14,15], P. Chawla[4,5], P. Demorest[13], M. Dobbs[4,5], E. Fonseca[4,5], U. Giri[16,17], D. C. Good[18], K. Gourdji[3], A. Josephy[4,5], A. Yu. Kirichenko[19,20], F. Kirsten[21], T. L. Landecker[22], D. Lang[16], T. J. W. Lazio[23], D. Z. Li[24,25], H.-H. Lin[25], J. D. Linford[13], K. Masui[26,27], J. Mena-Parra[26], A. Naidu[4,5], C. Ng[14], C. Patel[4,5], U.-L. Pen[14,16,25,28], Z. Pleunis[4,5], M. Rafiei-Ravandi[16], M. Rahman[14], A. Renard[14], P. Scholz[14,22], S. R. Siegel[4,5], K. M. Smith[16], I. H. Stairs[18], K. Vanderlinde[14,15] & A. V. Zwaniga[4,5]

Fast radio bursts (FRBs) are brief, bright, extragalactic radio flashes[1,2]. Their physical origin remains unknown, but dozens of possible models have been postulated[3]. Some FRB sources exhibit repeat bursts[4–7]. Although over a hundred FRB sources have been discovered[8], only four have been localized and associated with a host galaxy[9–12], and just one of these four is known to emit repeating FRBs[9]. The properties of the host galaxies, and the local environments of FRBs, could provide important clues about their physical origins. The first known repeating FRB, however, was localized to a low-metallicity, irregular dwarf galaxy, and the apparently non-repeating sources were localized to higher-metallicity, massive elliptical or star-forming galaxies, suggesting that perhaps the repeating and apparently non-repeating sources could have distinct physical origins. Here we report the precise localization of a second repeating FRB source[6], FRB 180916.J0158+65, to a star-forming region in a nearby (redshift 0.0337 ± 0.0002) massive spiral galaxy, whose properties and proximity distinguish it from all known hosts. The lack of both a comparably luminous persistent radio counterpart and a high Faraday rotation measure[6] further distinguish the local environment of FRB 180916.J0158+65 from that of the single previously localized repeating FRB source, FRB 121102. This suggests that repeating FRBs may have a wide range of luminosities, and originate from diverse host galaxies and local environments.

The CHIME/FRB Collaboration is discovering repeating FRB sources[5,6], which allows subsequent targeted observations using distributed radio telescope arrays to obtain precise interferometric localizations. CHIME/FRB discovered[6] the repeating source FRB 180916.J0158 + 65, which we observed with the very long baseline interferometry (VLBI) technique at a central frequency of 1.7 GHz and bandwidth of 128 MHz using eight radio telescopes of the European VLBI Network (EVN) for 5.5 h on 19 June 2019. As described in the Methods, we simultaneously recorded both EVN single-dish raw voltage data as well as high-time-resolution intensity data using the PSRIX data recorder[13] in filterbank mode at the 100-m Effelsberg telescope.

In a search of the PSRIX data, we detected four bursts from FRB 180916.J0158+65 with signal-to-noise ratios S/N between 9.5 and 46. The observed dispersion measures (DM) of the bursts are consistent with those previously reported[6] for this source. The burst properties, as derived from the PSRIX data, are listed in Extended Data Table 1.

[1]Joint Institute for VLBI ERIC (JIVE), Dwingeloo, The Netherlands. [2]ASTRON, Netherlands Institute for Radio Astronomy, Dwingeloo, The Netherlands. [3]Anton Pannekoek Institute for Astronomy, University of Amsterdam, Amsterdam, The Netherlands. [4]Department of Physics, McGill University, Montréal, Québec, Canada. [5]McGill Space Institute, McGill University, Montréal, Québec, Canada. [6]Max-Planck-Institut für Radioastronomie, Bonn, Germany. [7]Department of Astronomy and Owens Valley Radio Observatory, California Institute of Technology, Pasadena, CA, USA. [8]Department of Physics and Astronomy, West Virginia University, Morgantown, WV, USA. [9]Center for Gravitational Waves and Cosmology, West Virginia University, Chestnut Ridge Research Building, Morgantown, WV, USA. [10]School of Mathematics, Physics, and Statistics, University of Newcastle, Newcastle upon Tyne, UK. [11]Lane Department of Computer Science and Electrical Engineering, Morgantown, WV, USA. [12]Academia Sinica Institute of Astronomy and Astrophysics, Hilo, HI, USA. [13]National Radio Astronomy Observatory, Socorro, NM, USA. [14]Dunlap Institute for Astronomy and Astrophysics, University of Toronto, Toronto, Ontario, Canada. [15]Department of Astronomy and Astrophysics, University of Toronto, Toronto, Ontario, Canada. [16]Perimeter Institute for Theoretical Physics, Waterloo, Ontario, Canada. [17]Department of Physics and Astronomy, University of Waterloo, Waterloo, Ontario, Canada. [18]Department of Physics and Astronomy, University of British Columbia, Vancouver, British Columbia, Canada. [19]Instituto de Astronomía, Universidad Nacional Autónoma de México, Ensenada, Baja California, Mexico. [20]Ioffe Institute, St Petersburg, Russia. [21]Department of Space, Earth and Environment, Chalmers University of Technology, Onsala Space Observatory, Onsala, Sweden. [22]Dominion Radio Astrophysical Observatory, Herzberg Astronomy and Astrophysics Research Centre, National Research Council Canada, Penticton, British Columbia, Canada. [23]Jet Propulsion Laboratory, California Institute of Technology, Pasadena, CA, USA. [24]Department of Physics, University of Toronto, Toronto, Ontario, Canada. [25]Canadian Institute for Theoretical Astrophysics, Toronto, Ontario, Canada. [26]MIT Kavli Institute for Astrophysics and Space Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [27]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA, USA. [28]Canadian Institute for Advanced Research, CIFAR Program in Gravitation and Cosmology, Toronto, Ontario, Canada. *e-mail: J.W.T.Hessels@uva.nl

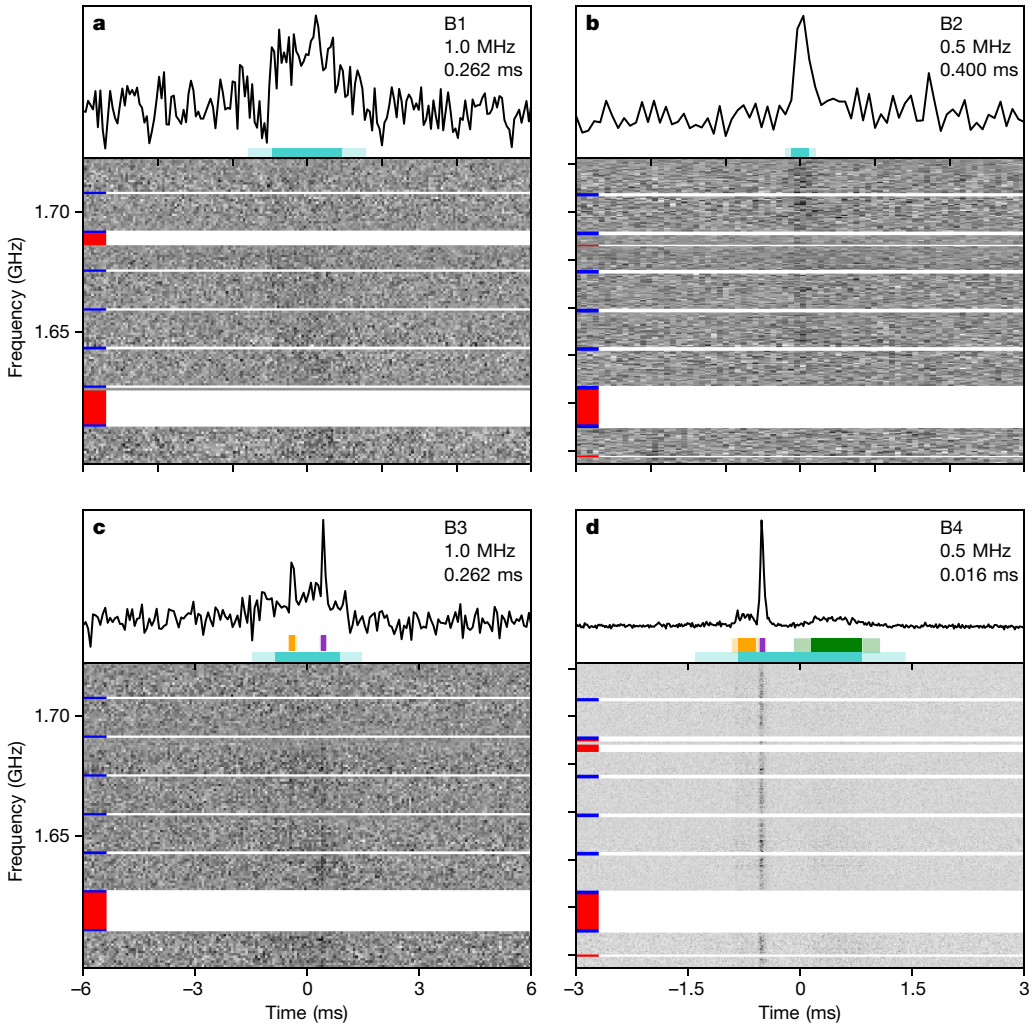


Fig. 1 | Burst detections in Effelsberg auto-correlation data. Band-averaged profiles and dynamic spectra of the four bursts, as detected using the coherently dedispersed Effelsberg auto-correlation data (**a**, **b**, **c** and **d**). Each burst was fitted with a Gaussian distribution to determine the full-width at half-maximum (FWHM) durations, which are represented by the dark cyan bars. The lighter cyan encloses the 2σ region. Bursts B3 and B4 (**c** and **d**) show sub-bursts; the orange, purple and green bars correspond to the FWHM of these sub-bursts, and the lighter bar (of each colour) encloses the 2σ region. Note the different time windows plotted: B1 and B3 (**a** and **c**) show 12 ms, whereas B2 and B4 (**b** and **d**) show 6 ms surrounding the burst peak. The solid white lines represent frequency channels that have been removed from the data due to either radio frequency interference (RFI) or subband edges, indicated by the red and blue markers, respectively. The burst data have been downsampled by various factors in time and frequency to optimize the visual representation. The RFI excision was done before downsampling. The time and frequency resolution used for plotting is shown in the top right of each panel.

No other dispersed single pulses of plausible astrophysical origin were found in this search, for DMs in the range 0–700 pc cm^{-3}. Using the EVN raw voltage data, we generated high-time-resolution (16-μs samples) Effelsberg auto-correlation data containing each burst. We used coherent dedispersion to mitigate the intra-channel smearing that dominated over the temporal resolution in the PSRIX data. By minimizing dispersion broadening, we properly resolve the burst structure, and find a best-fit DM = 348.76 ± 0.10 pc cm^{-3} using the brightest burst (Methods). We also detect brightness modulation with a characteristic frequency scale of 59 ± 13 kHz, which we interpret as scintillation imparted by the ionized interstellar medium of the Milky Way (Methods). The properties of the four bursts, as seen in the Effelsberg auto-correlations, are also shown in Extended Data Table 1. The frequency-averaged burst profiles and dedispersed dynamic spectra are shown in Fig. 1 and Extended Data Fig. 1 for the Effelsberg auto-correlation and PSRIX data, respectively.

We then used the EVN raw voltage data to create coherently dedispersed cross-correlations, also known as visibilities, at the times of the four bursts. Radio interferometric images with milli-arcsecond resolution were produced for each individual burst, using a single 0.7–2.7-ms integration, depending on the burst duration. Each image shows emission above at least seven times the root-mean-square (r.m.s.) noise level (Fig. 2; see Methods for a detailed explanation of the calibration and imaging process). The four burst images provide an average J2000 position for FRB 180916.J0158+65 of α = 01 h 58 min 00.75017 s ± 2.3 mas, δ = 65° 43′ 00.3152″ ± 2.3 mas. Visibilities with 2-s time resolution were also generated for the entire observation span. We used these to produce an image of the field around FRB 180916.J0158+65 in order to search for a persistent radio counterpart, like that seen in the case of FRB 121102[9,14]. No such emission is detected above a 3σ r.m.s. noise level of 30 μJy per beam. Independent observations using the Karl G. Jansky Very Large Array (VLA) at 1.6 GHz detect no coincident emission above a 3σ r.m.s. noise level of 18 μJy per beam (see Methods and Extended Data Fig. 2).

The precise EVN position shows that FRB 180916.J0158+65 is spatially coincident with a galaxy catalogued in the Sloan Digital Sky Survey[15] as SDSS J015800.28+654253.0. Given the maximum expected redshift of approximately 0.11, as inferred by the measured DM (Methods), we find that the probability for chance coincidence is less than 1% for any type of galaxy with mass greater than about 40% that of the FRB 121102

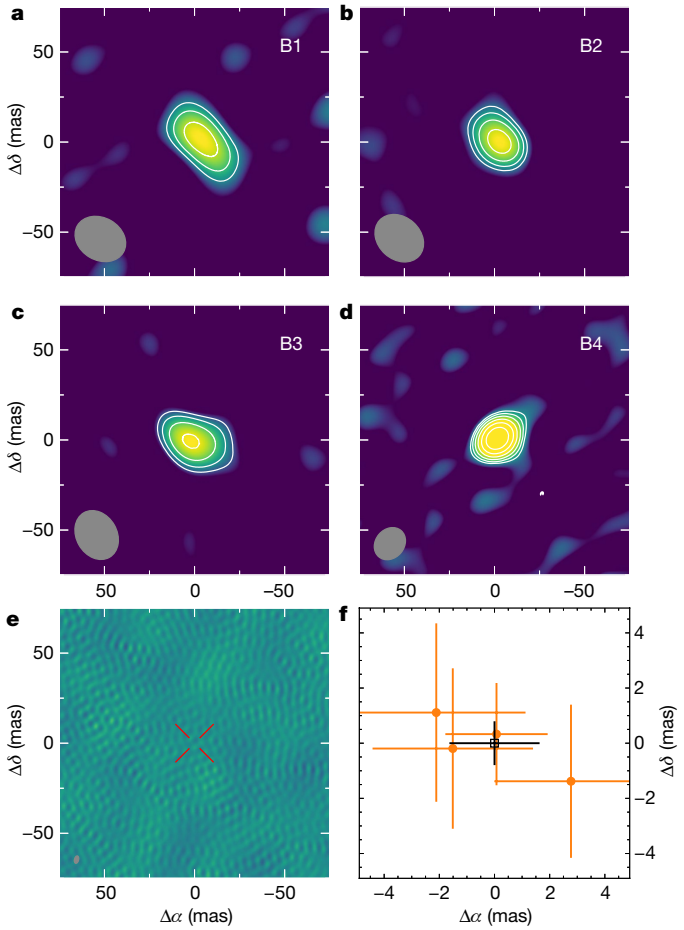


Fig. 2 | EVN images and burst positions. Each individual burst is imaged by using the EVN gated visibility data (**a**, **b**, **c** and **d**). **e**, Image using the full span of the EVN observation and 2-s integration time; no significant (>3σ) persistent radio emission is detected at the positions of the bursts (denoted by the red cross) and no signal above the 4σ level is detected anywhere within the full field. Contours start at three times the r.m.s. noise level of each image (75 mJy per beam, 65 mJy per beam, 43 mJy per beam and 100 mJy per beam and 9.7 μJy per beam, respectively) and increase by factors of $\sqrt{2}$. For all images, the synthesized beam is represented by the grey ellipse in the bottom-left corner. The Tianma station was only included in the derivation of the continuum image (**e**). **f**, The positions derived from each individual burst and their associated 1σ uncertainties (orange), with respect to the weighted-average burst position (black). All positions are referenced to right ascension α (J2000) = 01 h 58 min 00.75017 s, declination δ (J2000) = 65° 43′ 00.3152″.

host (Methods). Moreover, we thus conclude that the association of FRB 180916.J0158+65 to this particular galaxy is well established. FRB 180916.J0158+65 is close to the plane of the Milky Way, with Galactic longitude l = 129.7° and latitude b = 3.7°. Its low DM excess compared to the Milky Way contribution (Methods) brought into question whether it could possibly be a Galactic disk or halo object[6], but the host galaxy association shows that it is clearly extragalactic.

We used the 8-m Gemini-North telescope to characterize the morphology of the host galaxy and to measure a spectroscopic redshift (Methods). Deep optical imaging reveals that the host is a nearly face-on spiral galaxy (Fig. 3) with a total stellar mass of approximately 10^{10} times that of the Sun (Methods), which is comparable to the total stellar mass of the Milky Way. No other comparably large and bright galaxy is visible in the broader field covered by Gemini-North (Extended Data Fig. 3). The milli-arcsecond precision of the EVN localization shows that the FRB source is close to a bright feature in the r'-band approximately 7 arcsec (projected separation of roughly 4.7 kpc) from the core of the host galaxy.

With Gemini-North long-slit spectroscopy, we simultaneously targeted the host galaxy centre and the offset position of FRB 180916.J0158+65 (Extended Data Fig. 3). This revealed strong Hα emission, and several other spectral lines frequently associated with star formation (Fig. 3). By measuring optical line ratios, we confirm that the host is indeed a star-forming galaxy (Extended Data Fig. 5).

Comparing with the rest frequencies of the lines, we find a redshift z = 0.0337 ± 0.0002, which corresponds to a luminosity distance of 149.0 ± 0.9 Mpc (or an angular size distance of 139.4 ± 0.8 Mpc) using standard cosmological parameters[16]. FRB 180916.J0158+65 is thus the closest-known FRB source with a robust host galaxy identification and measured redshift. It is a factor of six closer than the repeater FRB 121102[17], and more than an order-of-magnitude closer than the (thus far) non-repeating sources FRB 180924[11], FRB 181112[12] and FRB 190523[10].

The optical spectrum at the location of FRB 180916.J0158+65 shows that both the associated star-forming clump and the host galaxy core are at the same redshift (Fig. 3). However, the spectrum at the location of FRB 180916.J0158+65 is dominated by emission from the clump, which is offset from the centre of the slit. Hence we have no independent direct estimate of the emission or dispersion measure at the position of FRB 180916.J0158+65 from the Gemini-North data. By considering the various modelled foreground contributions to DM, we estimate that the host contribution is less than approximately 70 pc cm^{-3}, and could be substantially smaller (Methods).

FRB 180916.J0158+65 is located at the apex of the v-shaped star-forming clump (Extended Data Fig. 4) with a relatively large[18] projected size of roughly 1.5 kpc and a star-formation rate of greater than about 0.016$M_\odot$ yr^{-1} (and a star-formation surface density of about $10^{-2}M_\odot$ yr^{-1} kpc^{-2}). The v-shaped star-forming clump is a remarkable feature of the galaxy, suggesting the possibility that the region has undergone an interaction that triggered the star formation, either between multiple star-forming regions or conceivably involving a putative dwarf satellite companion.

The spiral host galaxy of FRB 180916.J0158+65 contains over 100 times more stellar mass and has a metallicity five times higher (see Extended Data Fig. 5 and Methods) than the dwarf host galaxy of FRB 121102. The discovery of this host demonstrates that some FRB sources exist in galaxies more similar to our own Milky Way. Previously, it has been noted[14,17,19] that the host of FRB 121102 is similar to the type of low-metallicity galaxies with high specific star-formation rate that are associated with hydrogen-poor superluminous supernovae and long-duration gamma-ray bursts. In contrast, the FRB 180916.J0158+65 host is unlike such galaxies; this weakens the case for a general link between all repeating FRB sources and these extreme astrophysical explosions. In a search of the Open Supernova Catalog[20], we find no previous supernovae or gamma-ray bursts at the location of FRB 180916.J0158+65.

The proximity of FRB 180916.J0158+65 constrains the presence of any persistent radio counterpart to a luminosity (at a 3σ confidence level) of $\nu L_\nu < 1.3 \times 10^{36}$ erg s^{-1} from the continuum EVN data (sensitive to milli-arcsecond scales) and to $\nu L_\nu < 7.6 \times 10^{35}$ erg s^{-1} from the VLA data (sensitive to emission on arcsecond scales; see Methods). Compared to the persistent source associated with the repeating FRB 121102, this upper limit implies that any such source associated with FRB 180916.J0158+65 must be at least 400 times fainter than the one associated with FRB 121102[14]. The previously determined[6] Faraday rotation measure (RM) of FRB 180916.J0158+65, RM = −114.6 ± 0.6 rad m^{-2}, is three orders of magnitude lower than that of FRB 121102[21], where RM ≈ 10^5 rad m^{-2}. As previously suggested[6], we conclude that FRB 180916.J0158+65 is located in a much less extreme local environment compared with FRB 121102, and that the physical mechanism for FRB repetition does not depend on such conditions. Nonetheless, models originally proposed for FRB 121102[22,23]—in which the bursts originate from a young and rapidly rotating magnetar—could potentially still explain the observed properties of FRB 180916.J0158+65 by invoking an age of approximately 300 yr, which is ten times older than that proposed for FRB 121102 (see Methods for a brief comparison to existing models).

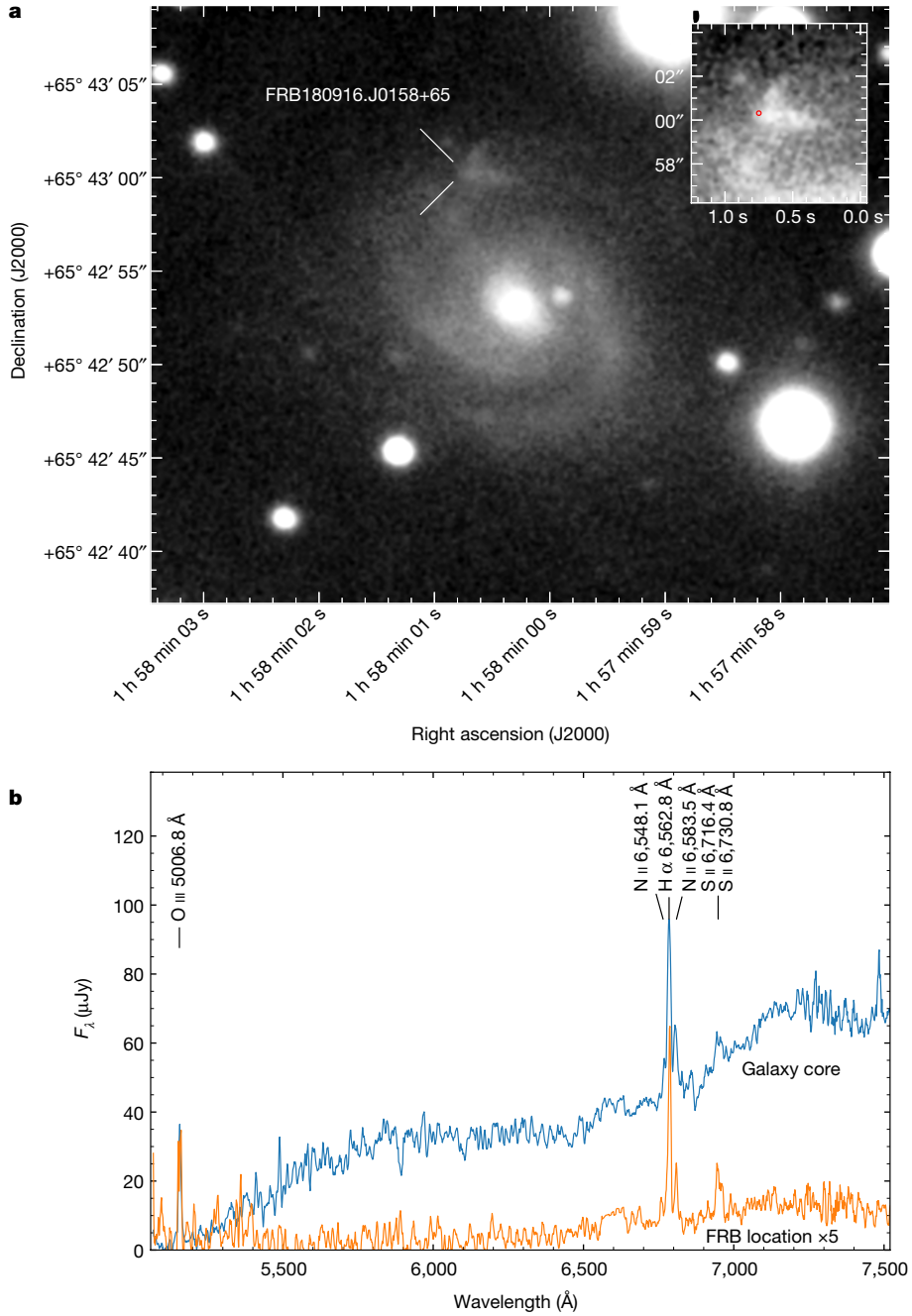


Fig. 3 | Gemini-North host galaxy image and optical spectrum. a, Image of the host galaxy using the *r*′ filter. The position of FRB 180916.J0158+65 is marked. The inset shows a higher-contrast zoom-in of the star-forming region containing FRB 180916.J0158+65 (marked by the red circle). The uncertainty in the position of FRB 180916.J0158+65 is smaller than the resolution of the image. **b**, Sky-subtracted spectrum extracted from a 5-arcsec aperture around the host galaxy core (blue) and a 2-arcsec aperture around the location of FRB 180916.J0158+65 (orange, scaled by a factor of five for clarity). Emission lines are identified along with their rest-frame wavelengths in air. Owing to the complicated shape of the galaxy, the flux densities, F_λ, have not been corrected for slit losses.

Although the host galaxies of FRB 180916.J0158+65 and FRB 121102 are markedly different, both sources are located near or within a star-forming region in the host galaxy. This contrasts with the elliptical host galaxies of FRB 180924[11] and FRB 190523[10], where there is comparatively little active star formation, but may be consistent with the star-forming galaxy of FRB 181112[12]. This diversity in hosts and local environments allows for the possibility that repeating and apparently non-repeating FRB sources have physically distinct origins. However, comparison of FRB event rates with those of proposed progenitors disfavours models that invoke cataclysmic explosions and suggests that a large fraction of sources must be capable of repeating[24]. The recent finding that FRB 171019 produces repeat bursts that are almost 600 times fainter than the originally discovered signal[7] underscores the fact that the detectability of repetition depends on instrumental sensitivity and source proximity. If FRB 180916.J0158+65 were at the distance of the other well localized FRBs, only a small fraction of its (brightest) bursts would be visible.

Furthermore, it has been proposed that a young magnetar origin for the bursts of as-yet non-repeating FRBs in non-star-forming regions is still viable as long as it is possible to form such sources through a variety

Table 1 | Burst properties

Burst	W_{tot} (ms)	W_{sub1} (ms)	W_{sub2} (ms)	W_{sub3} (ms)	Fluence (Jy ms)[a]	Spectral energy density (10^{28} erg Hz^{-1})[a]	$\Delta\alpha$ (mas)	$\Delta\delta$ (mas)	S_ν (Jy)[b]
B1	1.86 ± 0.13	–	–	–	0.72	1.90	−2.1 ± 3.2	1.1 ± 3.2	0.74 ± 0.08
B2	0.24 ± 0.02	–	–	–	0.20	0.53	−1.5 ± 2.9	−0.2 ± 2.9	0.66 ± 0.05
B3	1.72 ± 0.14	0.14 ± 0.02	0.12 ± 0.01	–	0.62	1.64	2.8 ± 2.8	−1.4 ± 2.8	0.50 ± 0.04
B4	1.66 ± 0.05[c]	0.24 ± 0.01	0.06 ± 0.0006	0.68 ± 0.03	2.53	6.68	0.1 ± 1.9	0.3 ± 1.9	2.3 ± 0.3

Widths, fluences and spectral energy densities are determined using the Effelsberg auto-correlation data of the bursts from FRB 180916.J0158+65, with a centre frequency of 1.7 GHz and dedispersed for a DM = 348.76 pc cm^{-3}. W_{tot} is the FWHM of the total burst envelope, and W_{subn} represents the FWHM of any sub-bursts, ordered according to their arrival time (the earliest sub-burst is labelled 'sub1'). The fluence is determined by integrating the burst profile over W_{tot} and is converted to physical units using the radiometer equation[34] (see Methods). The spectral energy density is estimated using the measured fluence and luminosity distance. The position offsets ($\Delta\alpha$, $\Delta\delta$) are referred to the average J2000 burst position of α = 01 h 58 min 00.75017 s, δ = +65° 43′ 00.3152″. The flux densities S_ν are measured by fitting a circular Gaussian to the EVN visibility data and are averages over the gate width.

[a]At 1.7-GHz reference frequency, and corresponding to the sum of all sub-bursts. A conservative fractional error of 30% is adopted for the derived fluences and energy densities. These are assumed to be isotropic.

[b]At 1.7-GHz reference frequency. The absolute flux scale may exhibit an additional ~15% uncertainty due to possible systematic gain calibration offsets.

[c]See Methods for how the total burst width of B4 was determined.

of channels, including direct stellar collapse, accretion-induced collapse, and through the merger of compact objects[25]. Ultimately, a larger number of precision localizations is needed before we can establish that multiple physical origins are required to explain the observed FRB phenomenon. There are now 11 repeating FRBs published[4–7], and with precision localizations it will be possible to establish whether repeating and apparently non-repeating FRB sources have demonstrably different environments.

FRBs have now been localized with luminosity distances that span approximately 150 Mpc to 4 Gpc. Estimating distance purely based on DM, it appears likely that there are FRBs that are even closer[26] or more distant[6,27] compared with this range. The four bursts presented here have isotropic-equivalent spectral energy densities as low as approximately 5×10^{27} erg Hz^{-1} (Table 1). Assuming similar beaming fractions, this makes them close to an order of magnitude less energetic than the weakest bursts seen from FRB 121102 to date[28], and between four to six orders-of-magnitude less energetic than FRB 180924[11] and FRB 190523[10]. Unless multiple models are invoked, a viable model for FRBs must address this large range of (apparent) energy outputs.

Comparing instead with pulsar emission, we note that the bursts from FRB 180916.J0158+65 are still at least a million times more energetic than the brightest giant pulses seen from the Crab pulsar—suggesting that they are not simply an exceptionally bright version of the known pulsar giant pulse phenomenon[29].

As previously suggested[6], the proximity of FRB 180916.J0158+65 is an advantage for multi-wavelength follow-up of the host galaxy and local environment. Whereas targeted observations of FRB 121102 have failed to detect either prompt or persistent optical, X-ray or gamma-ray counterparts[30–33], similar observations towards FRB 180916.J0158+65 may strongly constrain magnetar-based models, even in the event of non-detections. FRB 180916.J0158+65 is thus one of the most promising known sources for understanding the nature of FRBs.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1866-z.

1. Lorimer, D. R., Bailes, M., McLaughlin, M. A., Narkevic, D. J. & Crawford, F. A bright millisecond radio burst of extragalactic origin. *Science* **318**, 777–780 (2007).
2. Petroff, E., Hessels, J. W. T. & Lorimer, D. R. Fast radio bursts. *Astron. Astrophys. Rev.* **27**, 4 (2019).
3. Platts, E. et al. A living theory catalogue for fast radio bursts. *Phys. Rep.* **821**, 1–27 (2019).
4. Spitler, L. G. et al. A repeating fast radio burst. *Nature* **531**, 202–205 (2016).
5. CHIME/FRB Collaboration. A second source of repeating fast radio bursts. *Nature* **566**, 235–238 (2019).
6. The CHIME/FRB Collaboration. CHIME/FRB discovery of eight new repeating fast radio burst sources. *Astrophys. J.* **885**, L24 (2019).
7. Kumar, P. et al. Faint repetitions from a bright fast radio burst source. Preprint at https://arxiv.org/abs/1908.10026 (2019).
8. Petroff, E. et al. FRBCAT: The Fast Radio Burst Catalogue. *Publ. Astron. Soc. Aust.* **33**, e045 (2016).
9. Chatterjee, S. et al. A direct localization of a fast radio burst and its host. *Nature* **541**, 58–61 (2017).
10. Ravi, V. et al. A fast radio burst localized to a massive galaxy. *Nature* **572**, 352–354 (2019).
11. Bannister, K. W. et al. A single fast radio burst localized to a massive galaxy at cosmological distance. *Science* **365**, 565–570 (2019).
12. Prochaska, J. X. et al. The low density and magnetization of a massive galaxy halo exposed by a fast radio burst. *Science* **366**, 231–234 (2019).
13. Lazarus, P. et al. Prospects for high-precision pulsar timing with the new Effelsberg PSRIX backend. *Mon. Not. R. Astron. Soc.* **458**, 868–880 (2016).
14. Marcote, B. et al. The repeating fast radio burst FRB 121102 as seen on milliarcsecond angular scales. *Astrophys. J.* **834**, L8 (2017).
15. Alam, S. et al. The eleventh and twelfth data releases of the Sloan Digital Sky Survey: final data from SDSS-III. *Astrophys. J. Suppl.* **219**, 12 (2015).
16. Wright, E. L. A cosmology calculator for the world wide web. *Publ. Astron. Soc. Pacif.* **118**, 1711–1715 (2006).
17. Tendulkar, S. P. et al. The host galaxy and redshift of the repeating fast radio burst FRB 121102. *Astrophys. J.* **834**, L7 (2017).
18. Gusev, A. S. Hierarchy and size distribution function of star formation regions in the spiral galaxy NGC 628. *Mon. Not. R. Astron. Soc.* **442**, 3711–3721 (2014).
19. Metzger, B. D., Berger, E. & Margalit, B. Millisecond magnetar birth connects FRB 121102 to superluminous supernovae and long-duration gamma-ray bursts. *Astrophys. J.* **841**, 14 (2017).
20. Guillochon, J., Parrent, J., Kelley, L. Z. & Margutti, R. An open catalog for supernova data. *Astrophys. J.* **835**, 64 (2017).
21. Michilli, D. et al. An extreme magneto-ionic environment associated with the fast radio burst source FRB 121102. *Nature* **553**, 182–185 (2018).
22. Margalit, B. & Metzger, B. D. A concordance picture of FRB 121102 as a flaring magnetar embedded in a magnetized ion-electron wind nebula. *Astrophys. J.* **868**, L4 (2018).
23. Metzger, B. D., Margalit, B. & Sironi, L. Fast radio bursts as synchrotron maser emission from decelerating relativistic blast waves. *Mon. Not. R. Astron. Soc.* **485**, 4091–4106 (2019).
24. Ravi, V. The prevalence of repeating fast radio bursts. *Nat. Astron.* **3**, 928–931 (2019).
25. Margalit, B., Berger, E. & Metzger, B. D. Fast radio bursts from magnetars born in binary neutron star mergers and accretion induced collapse. *Astrophys. J.* **886**, 110 (2019).
26. Mahony, E. K. et al. A search for the host galaxy of FRB 171020. *Astrophys. J.* **867**, L10 (2018).
27. Bhandari, S. et al. The survey for pulsars and extragalactic radio bursts. II. New FRB discoveries and their follow-up. *Mon. Not. R. Astron. Soc.* **475**, 1427–1446 (2018).
28. Gourdji, K. et al. A sample of low-energy bursts from FRB 121102. *Astrophys. J.* **877**, L19 (2019).
29. Lyutikov, M. Fast radio bursts' emission mechanism: implication from localization. *Astrophys. J.* **838**, L13 (2017).
30. Scholz, P. et al. The repeating fast radio burst FRB 121102: multi-wavelength observations and additional bursts. *Astrophys. J.* **833**, 177 (2016).
31. Scholz, P. et al. Simultaneous X-ray, gamma-ray, and radio observations of the repeating fast radio burst FRB 121102. *Astrophys. J.* **846**, 80 (2017).
32. Hardy, L. K. et al. A search for optical bursts from the repeating fast radio burst FRB 121102. *Mon. Not. R. Astron. Soc.* **472**, 2800–2807 (2017).
33. MAGIC Collaboration. Constraining very-high-energy and optical emission from FRB 121102 with the MAGIC telescopes. *Mon. Not. R. Astron. Soc.* **481**, 2479–2486 (2018).
34. Cordes, J. M. & McLaughlin, M. A. Searches for fast radio transients. *Astrophys. J.* **596**, 1142–1154 (2003).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

A priori localization

The CHIME/FRB Collaboration discovered[6] the repeating source FRB 180916.J0158+65 and refined its position using the CHIME/FRB baseband mode. The automatic baseband triggering system of CHIME/FRB[35] captured[6] raw voltages for a burst from FRB 180916.J0158+65. The intensity of the signal in the region surrounding the original detection was then calculated by producing a grid of tightly packed tied-array beams. A refined position was obtained by fitting this intensity map with a simple model of the telescope beam. The technique will be described in more detail in a future paper (D.M. et al., manuscript in preparation). Since the analysis strategy was still in a preliminary stage at the time, we estimate a systematic uncertainty on this position of roughly 3 arcmin, based on a small sample of four known pulsars. CHIME/FRB baseband localizations are expected to have higher precision in the future.

European VLBI Network and Effelsberg observations

The EVN observed the field of FRB 180916.J0158+65 on 19 June 2019 for 5.5 h at a central frequency of 1.7 GHz and with a bandwidth of 128 MHz. The phase centre was placed at the position provided by the initial CHIME/FRB baseband localization: α = 01 h 57 min 43.2 s, δ = 65° 42′ 01.02″ (J2000 coordinates).

A total of eight dishes participated in the EVN observations: the 100-m Effelsberg, the 65-m Tianma, the 32-m Medicina, the 32-m Toruń, the 32-m Irbene, the 25-m × 38-m Jodrell Bank Mark II, the 25-m Onsala, and a single 25-m dish from the Westerbork array. The data were streamed in real time (e-EVN setup) to the EVN Software Correlator (SFXC[36]) at the Joint Institute for VLBI ERIC (JIVE) in Dwingeloo, The Netherlands. The real-time visibility data are comprised of eight 16-MHz subbands of 32 channels each, with full circular polarization products, and 2-s time averaging.

In parallel, we buffered the individual station raw voltage data in order to allow high-time-resolution correlations afterwards, at the times of any detected bursts. This method allows one to recover the signal from any position within the primary beam of the antennas (which have a full-width at half-maximum, FWHM, of roughly 7 arcmin in the case of the 100-m Effelsberg dish, and roughly 30 arcmin for the 25-m antennas), which would otherwise be smeared due to time and frequency averaging if it is more than tens of arcseconds away from the phase centre used in cross-correlation. We observed 3C454.3 and J0745+1011 as fringe-finders and bandpass calibrators. J0207+6246 was observed as phase calibrator (located 3.1° away from FRB 180916.J0158+65) in a phase-referencing cycle of 2 min on the calibrator and 5 min on the target, FRB 180916.J0158+65. We also observed J0140+6346 as a check source (3.1° away from the same phase calibrator) following the same phase-referencing cycle.

Simultaneously, we recorded high-time-resolution filterbank data at 1.7 GHz using the 100-m Effelsberg telescope and the PSRIX data recorder[13], which is designed for pulsar observations. We observed FRB 180916.J0158+65 for a total on-source time of 3.47 h in the frequency range 1597–1737 MHz. This total bandwidth was divided into 144 spectral channels of 0.98 MHz each. Of these, the bottom 48 channels were corrupted by RFI, probably from Iridium satellites. Consequently, these channels were removed from the data before beginning the analysis, giving a usable bandwidth of 93.75 MHz. The data were recorded with full Stokes information and a time resolution of 81.92 μs. Before the beginning of the observation, we performed a short test scan on the known pulsar PSR B2111+46, to verify the data integrity.

Interferometric data reduction

The EVN data were analysed using standard procedures within the Astronomical Image Processing System (AIPS[37]) and Difmap[38] software packages. A priori amplitude calibration was performed using the known gain curves and system temperature measurements recorded at each station during the observations. Poor data, primarily due to RFI, were flagged manually (less than about 10% of the total data). The remaining data were then fringe-fitted and bandpass calibrated using the fringe-finders and phase calibrator, which were imaged and self-calibrated to improve the final calibration of the data. The obtained solutions were transferred to FRB 180916.J0158+65 and J0140+6346 before creating the final images.

The check source J0140+6346 was used to estimate both the absolute astrometric uncertainty and the potential amplitude losses that could have been introduced due to the phase-referencing technique. Although the latter accounted for less than approximately 10% of the total flux density scale, we found a larger-than-expected offset (about 4 mas) of the centroid of the check source position with respect to its known coordinates from the International Celestial Reference Frame. The origin of this offset is explained by both the uncertainties associated with the phase-referencing technique[39–41] and the extended structure of J0207+6246, the phase calibrator. The core of this source shows multiple components that make the determination of the true position of the source ambiguous at the milli-arcsecond level. We corrected for this observed offset in all burst positions presented in this study, but an uncertainty on the final absolute positions of ±1.7 mas and ±2.1 mas in right ascension and declination, respectively, still remains owing to the ambiguity in the exact position of this reference source and its extension (the source is resolved on milli-arcsecond scales).

Search for FRB 180916.J0158+65 bursts

The PSRIX data were analysed in order to search for dispersed, millisecond-duration bursts, using the PRESTO[42] software package. The tool *rfifind* was used to identify frequency channels and time samples contaminated by RFI. We estimate that ~4% of the data were masked using *rfifind*. Incorporating the RFI mask, the data were then incoherently dedispersed to DMs in the range 0–700 pc cm^{-3} in steps of 1 pc cm^{-3} using *prepsubband*, generating dedispersed time series at each trial DM. To search for short-duration bursts, each dedispersed time series was convolved with boxcar functions of various widths, using *single_pulse_search.py*. Given the frequency resolution of the PSRIX data, the residual dispersive delay within each channel results in a temporal smearing of about 0.6 ms at the expected[6] DM of FRB 180916.J0158+65 (349 pc cm^{-3}). This intra-channel smearing dominates over the sampling time, causing a loss of sensitivity to extremely narrow bursts. We were sensitive to bursts with widths exceeding the intra-channel smearing time and up to 98.3 ms. A DM-versus-time plot (known as a single-pulse plot, generated by *single_pulse_search.py*) was created for each scan and inspected by eye for candidates. For each candidate identified, the dynamic spectrum was generated and inspected by eye to distinguish between astrophysical signals and terrestrial sources of RFI. During our 3.47-h on-source time, we found four bursts from FRB 180916.J0158+65 at its known DM. As a final check, each dedispersed time series was filtered for RFI in the search for candidates with peak $S/N > 7$, using an automated classifier[43,44] that returned only the same four bursts found in the initial manual search.

Using the time of arrival of each individual burst, one-second-duration Effelsberg auto-correlations were created for each burst, using the recorded EVN raw-voltage data. The auto-correlation data were generated with a higher temporal resolution (16-μs time bins) and frequency resolution (62.5-kHz channels) compared with the PSRIX data, and were coherently dedispersed to 350 pc cm^{-3}. Coherent dedispersion mitigates the effects of intra-channel smearing to less than 3 μs for DMs that are within a few units of 350 pc cm^{-3}. This provided us with the opportunity to better study the burst structure (limited only by our temporal resolution and the S/N), which is unresolved in the PSRIX data.

We identified the four bursts from FRB 180916.J0158+65 in the auto-correlations (Fig. 1 shows their band-averaged burst profiles and dynamic spectra). For comparison, we also show the profiles and dynamic spectra for the bursts in the PSRIX data (Extended Data Fig. 1).

Burst properties

Using the brightest burst, labelled B4, we determined a precise DM for FRB 180916.J0158+65. We used the PSRCHIVE[45] tool *pdmp* to determine the DM that corresponds to a peak in the *S*/*N*. Because B4 has a bright and narrow sub-burst, optimizing the *S*/*N* is probably also equivalent to optimizing the structure of the burst[46], in this case. We searched DMs of 350 ± 3 pc cm^{-3} in steps of 0.01 pc cm^{-3} using *pdmp*, and found the optimal DM to be 348.76 ± 0.10 pc cm^{-3} (consistent with the DM of 349.2 ± 0.3 pc cm^{-3} determined for this source by CHIME/FRB[6]). The bursts shown in both Fig. 1 and Extended Data Fig. 1 are dedispersed to this DM value.

Table 1 lists the physical properties of the bursts using the Effelsberg auto-correlations, dedispersed to 348.76 pc cm^{-3}. We measured the burst FWHM duration and peak time using a Gaussian fit[46]. The brightest burst, B4, shows a complicated structure (as can be seen in Fig. 1). The width of each individual sub-burst in B4 was determined using the same Gaussian fit as above. The total width of B4, however, was measured by defining the beginning of the burst as the peak time of the first sub-burst minus half W_{sub1}, and the end of the burst as the peak time of the last sub-burst plus half W_{sub3}.

To determine the fluence, we define a larger time window, determined by eye, than the quoted FWHM durations, which encompasses all of the burst flux. First, we produce a time series containing each burst by summing over all frequency channels. We define an off-pulse region, and use this to normalize the time series. We then convert the signal in each time bin to physical units by using the radiometer equation[34], and sum over the burst duration to estimate the fluence. It should be noted that B1 and B3 were downsampled by a factor of four for this analysis, resulting in a lower time resolution than the other bursts. For the radiometer equation, we take typical values for the system temperature $T_{\mathrm{sys}} \approx 20$ K and telescope gain $G \approx 1.54$ K Jy^{-1} for Effelsberg. For the fluence values quoted in Table 1, we provide a conservative fractional error of 30%. Also shown in Table 1 are the burst spectral energy densities[47], using the derived burst fluences and luminosity distance to FRB 180916.J0158+65. All four bursts are much dimmer than those previously detected[6] with CHIME/FRB, which is unsurprising given the higher sensitivity of the 100-m Effelsberg telescope.

Extended Data Table 1 shows the detection properties of the bursts, as seen in both the PSRIX data and the Effelsberg auto-correlations. For the PSRIX data, we quote the maximum *S*/*N* detected using PRESTO's *single_pulse_search.py* with an associated DM. The burst *S*/*N* for the auto-correlations is that measured at DM = 348.76 pc cm^{-3}, again using *single_pulse_search.py*. The burst width in Extended Data Table 1 is the boxcar width used in *single_pulse_search.py* resulting in the peak *S*/*N* value. The *S*/*N* of the four bursts, as discovered in the PSRIX data, are shown in Extended Data Fig. 6 as a function of time during our observation. The first three bursts appear clustered in time, occurring within 1.5 h of the beginning of our observation.

To study frequency-dependent brightness variations in the dominant sub-burst of B4—which are expected if there is scintillation—we conducted a standard auto-correlation analysis[48]. We generated the spectrum of the burst within the FWHM duration (W_{sub2} = 0.06 ms) using Effelsberg auto-correlations with a frequency resolution of 15.6 kHz. Spectral channels containing substantial RFI, along with those at the edges of the 16-MHz subbands, were set to zero in the spectrum before generating the auto-correlation function (ACF).

We calculated the ACF using the following expression:

$$\mathrm{ACF}(\Delta\nu) = \frac{\sum_i (S(\nu_i) - \bar{S})(S(\nu_i + \Delta\nu) - \bar{S})}{\sqrt{\sum_i (S(\nu_i) - \bar{S})^2}\sqrt{\sum_i (S(\nu_i + \Delta\nu) - \bar{S})^2}} \quad (1)$$

where ν is the frequency, $\Delta\nu$ is the frequency lag, and $\bar{S}$ is the mean of the spectrum. In this expression each summation is over the indices i which give non-zero values for both $S(\nu_i)$ and $S(\nu_i + \Delta\nu)$ (that is, we only include the 'good' data in the auto-correlation analysis). The ACF calculated using equation (1) is shown in Extended Data Fig. 7.

The central part of the ACF (for lags in the range −1.016 to +1.016 MHz) after removing the zero lag was then fitted with a Lorentzian function[49] of the form:

$$\frac{a}{x^2 + (\Delta\nu_{\mathrm{d}})^2} + b \quad (2)$$

using a least-squares fit, for free parameters: amplitude a; the decorrelation bandwidth (also known as the scintillation bandwidth) $\Delta\nu_{\mathrm{d}}$, defined[48] as the half-width at half-maximum of the ACF); and offset b. Extended Data Fig. 7 shows the fit to the ACF. The Lorentzian fit, with 127 degrees of freedom, returned a χ^2 of 105.2. We assume that the ACF between 3 MHz and 20 MHz lags contains no structure and represents the noise in the central profile of the ACF. We measure the standard deviation in this region and use this as our uncertainty for the calculation of χ^2. The P value of the fit is 0.92.

We measure the scintillation bandwidth to be 59 ± 13 kHz. This estimate is consistent with the NE2001 model[50] prediction of about 60 kHz along the line-of-sight at our observing frequency of 1.7 GHz. Using the simple relationship

$$2\pi\tau\Delta\nu_{\mathrm{d}} \approx 1 \quad (3)$$

we estimate a scattering timescale τ of about 2.7 μs, which we find is also similar to the NE2001 model prediction of around 2 μs along the line of sight[50]. This constraint on the scattering time is much tighter compared to CHIME/FRB measurements[6] of FRB 180916.J0158+65 bursts, even after accounting for the different observing frequencies and the typical ν^{-4} scaling of scattering time. CHIME/FRB nominally measured scattering times of approximately 2 ms for two FRB 180916.J0158+65 bursts, but these measurements are probably a reflection of burst morphology as opposed to genuine scattering effects (as was previously considered[6]).

High-time-resolution correlation and burst imaging

Coherently dedispersed visibilities for each burst were created by re-correlating the EVN data for only the time ranges corresponding to the burst durations. The exact time ranges were initially determined from the Effelsberg PSRIX data and then refined using coherently dedispersed Effelsberg auto-correlations. From the resulting auto-correlations we created pulse profiles and used these to set the optimal correlator gates applied to the interferometric data.

The a priori position of FRB 180916.J0158+65 provided at the time of the EVN observation had an uncertainty of about 3 arcmin (see above), making it infeasible to localize the bursts by directly imaging this region at milli-arcsecond resolution (which would require an image with around 10^{11} pixels). However, for strong signals, it is possible to estimate the position of the source from the geometric delays across the different telescope baselines. The differential geometric phase delay $\Delta\phi_{\mathrm{g}}$ of the source with respect to the phase centre (a priori pointing position) for a given baseline is given by[51,52]:

$$\Delta\phi_{\mathrm{g}} = 2\pi\nu(u\Delta\alpha\cos\delta + v\Delta\delta) \quad (4)$$

where ν is the observed frequency, $(\Delta\alpha, \Delta\delta)$ is the positional offset of the source with respect to the phase centre in right ascension (α) and declination (δ), respectively, and (u, v) are the coordinate offsets for the given baseline.

We used the strongest burst, B4, to estimate the offset of the source with respect to the phase centre. We then re-correlated the data at this refined J2000 position of α = 01 h 58 min 00.5 s, δ = 65°43′ 01.0″ (with an estimated uncertainty of a few arcseconds), which was around 2 arcmin from the a priori position—and thus consistent within the

estimated uncertainties. We applied the previously described EVN calibration to these burst data and imaged the individual bursts. All bursts were detected above a 7σ confidence level. Different imaging schemes were used during the imaging process in order to emphasize resolution (that is, contribution from the longest baselines) or sensitivity (that is, contribution from the core of the array). The final images were produced by using a natural weighting and excluding data from the Tianma station which, despite providing the highest resolution, exhibited potential calibration issues that could affect the absolute burst positions by a few milli-arcseconds—especially for such short integration times, where the uv coverage is poor. Table 1 lists the burst properties obtained from the EVN images, which are shown in Fig. 2. An image of the combined burst data was also produced. This final image is dominated by the emission of the strongest burst, B4, and thus we only consider the results from the four bursts individually.

The flux densities and positions were measured using Difmap and the Common Astronomy Software Applications (CASA) software packages by fitting a circular Gaussian distribution component to the detected emission in the uv plane, which is expected to be a robust method against station gain calibration uncertainties[53]. The position of each burst was measured independently and then we determined a weighted average J2000 position for the source of α = 01 h 58 min 00.75017s ± 2.3 mas, δ = 65° 43′ 00.3152″ ± 2.3 mas, where we used the fluence of each burst as weights. This is within 1.7 arcsec of the position determined via mapping the geometric delays. The final positional uncertainty reflects the statistical uncertainties from the individual position measurements (about 1 mas), the uncertainties in the absolute International Celestial Reference Frame position of the phase calibrator (J0207+6246; 0.15 mas) and check source (J0140+6346; 0.1 mas), and the systematic uncertainty associated with the phase-referencing technique and the extended structure of the check source and phase calibrator (around 2 mas), as explained previously. Uncertainties were added in quadrature.

Finally, we searched for continuum radio counterparts of FRB 180916.J0158+65 by imaging the full interferometric dataset with a field-of-view of 2 arcsec × 2 arcsec centred at the position of the bursts (and thus also covering the full extent of the nearby star-forming region). We detected no emission at the position of FRB 180916.J0158+65 as well as no significant (>4.5σ) compact emission anywhere in the field above the r.m.s. noise level of 10 μJy per beam (19 μJy per beam when excluding the data from the Tianma station). We also searched for possible emission from the core of the galaxy, but found no signal above the 6σ r.m.s. noise level in a 4 arcsec × 4 arcsec area.

VLA observations

In June 2019, we used the VLA to perform a quasi-simultaneous search for millisecond transients and faint persistent emission associated with FRB 180916.J0158+65. The VLA observed from 1 GHz to 2 GHz with baseline lengths up to 11 km, giving a synthesized beam size of roughly 4 arcsec. VLA visibilities were sampled at 5-ms time resolution and searched in real time by the Realfast system[54]; the same data were integrated at 3-s time resolution and calibrated with the standard CASA calibration pipeline. In a total of 14 h of observing (with on-target efficiency of 80%), we found no millisecond transients to an 8σ limit of approximately 50 mJy per beam in 5 ms or, equivalently, a fluence limit of 0.25 Jy ms. No VLA observing was coincident with known burst times detected by the EVN. We concatenated all 14 h of data, applied an iteration of phase self-calibration using relatively bright background sources in the field of view of the antennas, and imaged 600 MHz of the bandwidth centred near 1.6 GHz to search for persistent radio emission on arcsecond scales. We confirmed that the astrometric accuracy of the deep image is good to better than 1 arcsec by associating seven bright field radio sources to the catalogued location in the NRAO VLA Sky Survey[55]. No persistent emission is detected at the position of FRB 180916.J0158+65 brighter than a 3σ limit of 18 μJy per beam. A 25-μJy source is located at a nearby J2000 position of α = 01 h 58 min 00.9 s, δ = 65° 43′ 07″ (see Extended Data Fig. 2), but this position is offset by approximately 7 arcsec from the location of FRB 180916.J0158+65, which is much more than the astrometric precision, estimated to be 0.4 arcsec (the VLA observed in its B configuration at the time). It is thus very unlikely to be associated with FRB 180916.J0158+65 or the host galaxy.

Optical observations and host galaxy

The EVN localization of FRB 180916.J0158+65 places it at the outskirts of SDSS J015800.28+654253.0, an apparent Sb spiral galaxy (possibly with a faint bar) with an estimated[15] photometric redshift $z_{\rm phot}$ = 0.07 ± 0.05.

We obtained imaging and spectroscopic observations of the host galaxy using the GMOS spectrograph on the Gemini-North telescope between July and September 2019 (programme ID GN-2019A-DD-110). We acquired 2 × 900-s exposures with each of the g' and r' filters. The images were processed with the Gemini IRAF/pyraf package, and sources were extracted using SExtractor[56]. The images were de-warped and astrometrically corrected by matching the point source positions to those from the Gaia Data Release 2 (DR2) Catalogue[57,58]. The matching precision was 17–20 mas r.m.s. for each image. The seeing was measured to be 0.99 and 0.78 arcsec in the g' and r' images, respectively.

Figure 3 shows the r' image of the galaxy, along with the position of FRB 180916.J0158+65. The galaxy is nearly face-on, with spiral arm structures visible. FRB 180916.J0158+65 appears to be co-located with one of the outer arms of the galaxy, near a bright clump in the r' image (see Extended Data Fig. 4 for a more detailed zoom on this region). We attribute this emission to that of Hα, which is bright in emission in the GMOS spectra and, at the redshift of the host galaxy, falls within the r' image. The total stellar mass of the galaxy has been estimated to be about 10^{10} times the mass of the Sun by using the WISE W1 and W2 colours[59], which provide a more robust estimation given that they are less affected by extinction.

We performed long-slit optical spectroscopy, simultaneously targeting the galaxy core and the FRB location, in order to measure the redshift and spectral properties of the galaxy and the location of FRB 180916.J0158+65. We acquired 4 × 900-s spectroscopic observations with a 1.5-arcsec slit and the R400 grating—along with corresponding flat fields, copper–argon arcs, and observations of BD +28 4211 to correct for sensitivity and extinction. As shown in Extended Data Fig. 4, the slit was rotated to align with the host galaxy centre and FRB position.

The spectroscopic data were processed with the Gemini IRAF/pyraf package. The FWHM of the emission lines from the sky was measured to be 7–8 Å, matching the theoretical resolving power of $R \approx 640$. The wavelength calibration with copper–argon emission lines had a precision of 0.7–0.8 Å. However, the lack of arc lines towards the bluer edge reduces the accuracy of the wavelength solution towards 5,200 Å.

We extracted the spectrum within 5 arcsec of the galaxy centre as well as 2 arcsec around the location of FRB 180916.J0158+65 itself. We used the trace of the galaxy centre, offset by 7.7 arcsec to trace the location of FRB 180916.J0158+65. Figure 3 shows the calibrated spectra at each location separately. We identify redshifted Hα, N II, S II and O III lines with an average redshift of z = 0.033 ± 0.001 if all lines are included and z = 0.0337 ± 0.0002 when excluding the O III lines that are at the blue edge of the charge-coupled device (CCD)—and which hence may have reduced wavelength calibration accuracy. Extended Data Table 2 shows the line centroids, widths and fluxes. The spectrum at the location of FRB 180916.J0158+65 is dominated by emission lines with little continuum emission. The O III 5,007 Å line is detected at the FRB location but was corrupted by a cosmic-ray hit. As the slit was positioned at the edge of the nearby star-forming clump, we observe a non-uniform slit illumination pattern and the line centroids, especially for Hα, are shifted towards the redder wavelengths by about 1 Å. Nevertheless, we can confirm that both the clump and the galaxy are at the same distance.

The flux obtained for the Hα emission at the position of FRB 180916.J0158+65 (which is dominated by the star-forming clump) implies a

luminosity of $L_{H\alpha} \approx (2.0 \pm 0.1) \times 10^{39}$ erg s^{-1}, which corresponds to a star-formation rate[60] of about $0.016 M_{\odot}$ yr^{-1}. These values were extracted from a region of 1.5 arcsec × 2 arcsec around the FRB, corresponding to 1.5 kpc^2, implying a star-formation surface density of about $10^{-2} M_{\odot}$ yr^{-1} kpc^{-2}. Note that these values can be strongly affected by extinction, which is estimated to be $E(B-V) \approx 1.01$.

From the Hα, N II and S II line fluxes, we estimate the metallicity of the galaxy[61] to be 12 + log(O/H) = 8.82, close to solar neighbourhood metallicity, and five times higher when compared to the host of FRB 121102 (<8.1 on the same metallicity scale[17]).

Chance alignment probability with host galaxy

The projected position of FRB 180916.J0158+65 is coincident with a star-forming clump in the spiral galaxy SDSS J015800.28+654253.0. To confirm an association with the galaxy we have estimated the probability of chance coincidence between the two sources.

We have considered a uniform density of galaxies across the sky. Assuming a Poisson distribution of galaxies in the relevant co-moving volume, we estimate the number density of typical galaxies brighter than $M_B = -16$ (that is, galaxies with masses greater than about 40% of the dwarf galaxy host[17] of FRB 121102) based on the all-galaxy luminosity function[62,63], and their typical size distribution based on the galaxies reported in SDSS DR7[64]. Then, we calculate the chance coincidence probability for a source to be co-located within twice the median half-light radius of any galaxy as a function of redshift.

At the measured redshift of $z \approx 0.0337$, the chance coincidence probability $P \ll 0.1\%$. However, we may consider the probability up to a redshift of approximately 0.11, the maximum redshift considered for FRB 180916.J0158+65 based on the observed DM[6]. Even in this case $P < 1\%$ (see Extended Data Fig. 8). The association between FRB 180916.J0158+65 and the galaxy is thus statistically robust.

DM budget

FRB 180916.J0158+65 is close to the Galactic plane, with longitude $l = 129.7°$ and latitude $b = 3.7°$. The total measured DM of FRB 180916.J0158+65 is $DM_{tot} = 348.76$ pc cm^{-3}, and is expected to be the sum of contributions from the Milky Way disk (DM_{MW}) and halo (DM_{halo}), the intergalactic medium (DM_{IGM}), and host galaxy (DM_{host}; including both the ISM and halo of the host). For the line-of-sight to FRB 180916.J0158+65, DM_{MW} is predicted to be 199 pc cm^{-3} and 325 pc cm^{-3} by the NE2001 and YMW16 Galactic electron density models, respectively[50,65]. DM_{halo} is estimated to be 50–80 pc cm^{-3} or approximately 45 pc cm^{-3}, according to two available models[12,66]. Given the range of model predictions, one can ascribe up to roughly 100 pc cm^{-3} to DM_{IGM} and DM_{host}. For redshift $z = 0.0337$, DM_{IGM} is expected to be[67] approximately 34 pc cm^{-3}—although this estimate is highly dependent on line of sight for low redshift[68]. Overall, after subtracting these various contributions, DM_{host} is probably less than 70 pc cm^{-3}, and could be much smaller. The YMW16 model over-predicts DM_{MW} in this direction (unless the DM_{halo}, DM_{IGM} and DM_{host} contributions are all much smaller than expected), and this illustrates that FRB DMs will help constrain Galactic electron density models. Conversely, because the line of sight to FRB 180916.J0158+65 passes through the Galactic plane, with a large but uncertain Galactic DM contribution, it is unlikely to be a good probe for refining the DM–z relationship in the IGM[67].

Comparison with the other localized repeating FRB

The physical origin of the FRB phenomenon remains unclear, and many of the proposed models[3] remain consistent with the observational facts we present here for FRB 180916.J0158+65. We now briefly consider the nature of FRB 180916.J0158+65 by comparison to two of the families of models proposed for FRB 121102, which is the FRB source studied in most detail to date, and the only other well localized repeater.

One of the models to explain the associated persistent radio counterpart to FRB 121102 proposes the existence of a 20–50-year-old supernova remnant and nebula powered by a young flaring magnetar[22,23,69,70]. The absence of a similarly bright radio nebula associated with FRB 180916.J0158+65 could be explained, in this model, as an older system whose emission has already faded. Assuming a similar model[22], the upper limits on persistent radio emission associated with FRB 180916.J0158+ 65 are consistent with a system that is at least about five times older than FRB 121102, and thus at least 200–500 yr old. Additionally, the same model also predicts a characteristic decay of the RM with time. An RM as large as the one observed towards FRB 121102[21] would drop to the RM observed[6] towards FRB 180916.J0158+65 at a source age of about 300 yr. However, while the results we present here are conceivably consistent with the young magnetar scenario for repeating FRBs, we find no new support for this scenario.

Other models, inspired originally by FRB 121102, invoke interaction between a nearby strong plasma stream and a neutron star's magnetosphere[71]. In the case of FRB 121102, the large[21] RM and the observational similarity[14] of its compact, persistent radio source to low-luminosity active galactic nuclei suggest that the bursts could originate from a massive black hole, or a neutron star in the near vicinity. An accreting massive black hole can provide a plasma stream that could interact with a neutron star magnetosphere[72]. The absence of a persistent radio counterpart to FRB 180916.J0158+65, and its location in the spiral arm of a morphologically well defined galaxy, make such a model arguably less likely in the case of FRB 180916.J0158+65. For a black hole with a similar mass to that of the one considered for FRB 121102 ($10^5 M_{\odot}$ to $10^6 M_{\odot}$), our upper limits would imply a very low accretion rate of about $10^{-7} L_{Edd}$ where L_{Edd} is the Eddington luminosity. Nonetheless, this accretion rate is still compatible with the low end observed in low-luminosity active galactic nuclei[76]. The presence of a massive black hole at the apex of the star-forming clump is thus not excluded, although rather unlikely. Nonetheless, in an interacting model[71], the plasma stream can also have a different origin, and it is conceivable that FRB 180916.J0158+65 is a neutron star interacting with, for example, the jet of a stellar-mass black hole.

The proximity of FRB 180916.J0158+65 will aid in performing deep multi-wavelength observations, which may help to discriminate between the various scenarios outlined above, while potentially also suggesting new ones.

Data availability

The datasets generated from the EVN observations and analysed in this study are available at the Public EVN Data Archive (http://www.jive.eu/select-experiment) under the experiment code EM135C. The datasets generated from the Gemini observations are available at the Public Gemini Observatory Archive (https://archive.gemini.edu) with Program ID GN-2019A-DD-110.

Code availability

The codes used to analyse the data are available at the following sites: AIPS (http://www.aips.nrao.edu/index.shtml), CASA (https://casa.nrao.edu), Difmap (ftp://ftp.astro.caltech.edu/pub/difmap/difmap.html), IRAF (http://ast.noao.edu/data/software), PRESTO (https://github.com/scottransom/presto), and PSRCHIVE (http://psrchive.sourceforge.net). This research made use of APLpy, an open-source plotting package for Python hosted at http://aplpy.github.com, Astropy, a community-developed core Python package for Astronomy[77], and Matplotlib[78].

35. CHIME/FRB Collaboration. The CHIME fast radio burst project: system overview. *Astrophys. J.* **863**, 48 (2018).
36. Keimpema, A. et al. The SFXC software correlator for very long baseline interferometry: algorithms and implementation. *Exp. Astron.* **39**, 259–279 (2015).
37. Greisen, E. W. AIPS, the VLA, and the VLBA. In *Information Handling in Astronomy. Historical Vistas* (ed. Heck, A.) Vol. 285, 109 (Astrophysics and Space Science Library, 2003).
38. Shepherd, M. C., Pearson, T. J. & Taylor, G. B. DIFMAP: an interactive program for synthesis imaging. *Bull. Am. Astron. Soc.* **26**, 987–989 (1994).

39. Chatterjee, S. et al. Pulsar parallaxes at 5 GHz with the Very Long Baseline Array. *Astrophys. J.* **604**, 339–345 (2004).
40. Pradel, N., Charlot, P. & Lestrade, J. F. Astrometric accuracy of phase-referenced observations with the VLBA and EVN. *Astron. Astrophys.* **452**, 1099–1106 (2006).
41. Kirsten, F., Vlemmings, W., Campbell, R. M., Kramer, M. & Chatterjee, S. Revisiting the birth locations of pulsars B1929+10, B2020+28, and B2021+51. *Astron. Astrophys.* **577**, A111 (2015).
42. Ransom, S. M. *New Search Techniques for Binary Pulsars*. PhD thesis, Harvard Univ. https://ui.adsabs.harvard.edu/abs/2001PhDT.....123R/abstract (2001).
43. Michilli, D. et al. Single-pulse classifier for the LOFAR tied-array all-sky survey. *Mon. Not. R. Astron. Soc.* **480**, 3457–3467 (2018).
44. Michilli, D. & Hessels, J. W. T. SpS: Single-pulse Searcher. *Astrophys. Source Code Library* 1806. 013 (2018).
45. Hotan, A. W., van Straten, W. & Manchester, R. N. PSRCHIVE and PSRFITS: an open approach to radio pulsar data storage and analysis. *Publ. Astron. Soc. Aust.* **21**, 302–309 (2004).
46. Hessels, J. W. T. et al. FRB 121102 bursts show complex time-frequency structure. *Astrophys. J.* **876**, L23 (2019).
47. Law, C. J. et al. A multi-telescope campaign on FRB 121102: implications for the FRB population. *Astrophys. J.* **850**, 76 (2017).
48. Cordes, J. M., Weisberg, J. M. & Boriakoff, V. Small-scale electron density turbulence in the interstellar medium. *Astrophys. J.* **288**, 221–247 (1985).
49. Rickett, B. J. Radio propagation through the turbulent interstellar plasma. *Annu. Rev. Astron. Astrophys.* **28**, 561–605 (1990).
50. Cordes, J. M. & Lazio, T. J. W. NE2001.I. A new model for the galactic distribution of free electrons and its fluctuations. Preprint at https://arxiv.org/abs/astro-ph/0207156 (2002).
51. Fomalont, E. B. & Perley, R. A. Calibration and editing. In *Synthesis Imaging in Radio Astronomy II* (eds Taylor, G. B., Carilli, C. L. & Perley, R. A.) Vol. 180, 79 (Astronomical Society of the Pacific Conference Series, 1999).
52. Thompson, A. R. Fundamentals of Radio Interferometry. In *Synthesis Imaging in Radio Astronomy II* (eds Taylor, G. B., Carilli, C. L. & Perley, R. A.) Vol. 180, 11 (Astronomical Society of the Pacific Conference Series, 1999).
53. Natarajan, I. et al. Resolving the blazar CGRaBS J0809+5341 in the presence of telescope systematics. *Mon. Not. R. Astron. Soc.* **464**, 4306–4317 (2017).
54. Law, C. J. et al. realfast: real-time, commensal fast transient surveys with the Very Large Array. *Astrophys. J. Suppl.* **236**, 8 (2018).
55. Condon, J. J. et al. The NRAO VLA sky survey. *Astron. J.* **115**, 1693–1716 (1998).
56. Bertin, E. & Arnouts, S. SExtractor: software for source extraction. *Astron. Astrophys. Suppl. Ser.* **117**, 393–404 (1996).
57. Gaia Collaboration. Gaia Data Release 1. Summary of the astrometric, photometric, and survey properties. *Astron. Astrophys.* **595**, A2 (2016).
58. Gaia Collaboration. Gaia Data Release 2. Summary of the contents and survey properties. *Astron. Astrophys.* **616**, A1 (2018).
59. Jarrett, T. H. et al. Galaxy and Mass Assembly (GAMA): exploring the WISE web in G12. *Astrophys. J.* **836**, 182 (2017).
60. Kennicutt, J., Robert, C., Tamblyn, P. & Congdon, C. E. Past and future star formation in disk galaxies. *Astrophys. J.* **435**, 22 (1994).
61. Dopita, M. A., Kewley, L. J., Sutherland, R. S. & Nicholls, D. C. Chemical abundances in high-redshift galaxies: a powerful new emission line diagnostic. *Astrophys. Space Sci.* **361**, 61 (2016).
62. Faber, S. M. et al. Galaxy luminosity functions to $z \sim 1$ from DEEP2 and COMBO-17: implications for red galaxy formation. *Astrophys. J.* **665**, 265–294 (2007).
63. Blanton, M. R. et al. The galaxy luminosity function and luminosity density at redshift $z = 0.1$. *Astrophys. J.* **592**, 819–838 (2003).
64. Zhang, Y.-C. & Yang, X.-H. Size distribution of galaxies in SDSS DR7: weak dependence on halo environment. *Res. Astron. Astrophys.* **19**, 006 (2019).
65. Yao, J. M., Manchester, R. N. & Wang, N. A new electron-density model for estimation of pulsar and FRB Distances. *Astrophys. J.* **835**, 29 (2017).
66. Yamasaki, S. & Totani, T. The galactic halo contribution to the dispersion measure of extragalactic fast radio bursts. Preprint at https://arxiv.org/abs/1909.00849 (2019).
67. Inoue, S. Probing the cosmic reionization history and local environment of gamma-ray bursts through radio dispersion. *Mon. Not. R. Astron. Soc.* **348**, 999–1008 (2004).
68. Li, Y., Zhang, B., Nagamine, K. & Shi, J. The FRB 121102 host is atypical among nearby fast radio bursts. *Astrophys. J.* **884**, L26 (2019).
69. Lyubarsky, Y. A model for fast extragalactic radio bursts. *Mon. Not. R. Astron. Soc.* **442**, L9–L13 (2014).
70. Beloborodov, A. M. A flaring magnetar in FRB 121102? *Astrophys. J.* **843**, L26 (2017).
71. Zhang, B. A "cosmic comb" model of fast radio bursts. *Astrophys. J.* **836**, L32 (2017).
72. Zhang, B. FRB 121102: a repeatedly combed neutron star by a nearby low-luminosity accreting supermassive black hole. *Astrophys. J.* **854**, L21 (2018).
73. Kewley, L. J., Dopita, M. A., Sutherland, R. S., Heisler, C. A. & Trevena, J. Theoretical modeling of starburst galaxies. *Astrophys. J.* **556**, 121–140 (2001).
74. Kewley, L. J. & Dopita, M. A. Using strong lines to estimate abundances in extragalactic H II regions and starburst galaxies. *Astrophys. J. Suppl.* **142**, 35–52 (2002).
75. Kauffmann, G. et al. The host galaxies of active galactic nuclei. *Mon. Not. R. Astron. Soc.* **346**, 1055–1077 (2003).
76. Loewenstein, M., Mushotzky, R. F., Angelini, L., Arnaud, K. A. & Quataert, E. Chandra limits on X-ray emission associated with the supermassive black holes in three giant elliptical galaxies. *Astrophys. J.* **555**, L21–L24 (2001).
77. Astropy Collaboration. Astropy: a community Python package for astronomy. *Astron. Astrophys.* **558**, A33 (2013).
78. Hunter, J. D. Matplotlib: a 2D graphics environment. *Comput. Sci. Eng.* **9**, 90–95 (2007).

Acknowledgements We thank W. J. G. de Blok, L. Connor, N. Maddox, E. Petroff, H. Vedantham and J. Weisberg for discussions. The European VLBI Network is a joint facility of independent European, African, Asian, and North American radio astronomy institutes. Scientific results from data presented in this publication are derived from the following EVN project code: EM135. This work was also based on simultaneous EVN and PSRIX data recording observations with the 100-m telescope of the Max-Planck-Institut für Radioastronomie at Effelsberg, and we thank the local staff for this arrangement. Our work is also based on observations obtained at the Gemini Observatory (programme DT-2019A-135), which is operated by the Association of Universities for Research in Astronomy, Inc., under a cooperative agreement with the NSF on behalf of the Gemini partnership: the National Science Foundation (United States), the National Research Council (Canada), CONICYT (Chile), Ministerio de Ciencia, Tecnología e Innovación Productiva (Argentina), and Ministério da Ciência, Tecnologia e Inovação (Brazil). B.M. acknowledges support from the Spanish Ministerio de Economía y Competitividad (MINECO) under grants AYA2016-76012-C3-1-P and MDM-2014-0369 of ICCUB (Unidad de Excelencia "María de Maeztu"). J.W.T.H. acknowledges funding from an NWO Vidi fellowship and from the European Research Council under the European Union's Seventh Framework Programme (FP/2007-2013)/ERC Starting Grant agreement number 337062 ("DRAGNET"). M.B. is supported by an FRQNT Doctoral Research Award, Physics Department Excellence Award and a Mitacs Globalink Graduate Fellowship. R.K. is supported by ERC synergy grant number 610058 ("BlackHoleCam"). V.M.K. holds the Lorne Trottier Chair in Astrophysics and Cosmology, a Canada Research Chair and the R. Howard Webster Foundation Fellowship of CIFAR. V.M.K. receives support from an NSERC Discovery Grant and Herzberg Award, and from the FRQNT Centre de Recherche en Astrophysique du Québec. C.J.L. acknowledges support from NSF grant 1611606. D.M. is a Banting Fellow. K.A. acknowledges support from NSF grant AAG-1714897. B.A. is supported by a Chalk-Rowles Fellowship. A.M.A. acknowledges funding from an NWO Veni fellowship. S.B.-S. acknowledges support from NSF grant AAG-1714897. F.K. thanks the Swedish Research Council. U.-L.P. receives support from the Ontario Research Fund Research Excellence Program (ORF-RE), NSERC, the Simons Foundation, Thoth Technology Inc., and the Alexander von Humboldt Foundation. Z.P. is supported by a Schulich Graduate Fellowship. P.S. is a Dunlap Fellow and an NSERC Postdoctoral Fellow. The Dunlap Institute is funded through an endowment established by the David Dunlap family and the University of Toronto. K.M.S. is supported by an NSERC Discovery Grant, an Ontario Early Researcher Award, and a CIFAR fellowship. Part of this research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under a contract with the National Aeronautics and Space Administration. The NANOGrav project receives support from National Science Foundation (NSF) Physics Frontiers Center award number 1430284. FRB research at the University of British Columbia (UBC) is supported by an NSERC Discovery Grant and by the Canadian Institute for Advanced Research. The CHIME/FRB baseband system is funded in part by a Canada Foundation for Innovation John R. Evans Leaders Fund award to I.H.S. The National Radio Astronomy Observatory (NRAO) is a facility of the National Science Foundation operated under cooperative agreement by Associated Universities, Inc. The Astronomical Image Processing System (AIPS) is a software package produced and maintained by NRAO. The Common Astronomy Software Applications (CASA) package is software produced and maintained by NRAO.

Author contributions B.M. is the Principal Investigator of the EVN observing programme and led the radio imaging analysis of those data. K.N. discovered the radio bursts and led the time-domain analysis. J.W.T.H. guided the time-domain analysis and made major contributions to the writing of the manuscript. S.P.T. and C.G.B. led the optical imaging and spectroscopic analyses. Z.P. performed an independent analysis of the EVN data, and in previous years pioneered the development of the EVN's capabilities for fast transient research. A.K. created the software used to correlate the EVN baseband data. M.B. analysed the chance coincidence probability. R.K. arranged the Effelsberg PSRIX observations. C.J.L. analysed the VLA imaging data. D.M. determined the CHIME/FRB baseband position. V.M.K. had an important coordination role, enabling these results to be gathered. All other co-authors contributed to the CHIME/FRB discovery of the source or the interpretation of the analysis results and the final version of the manuscript.

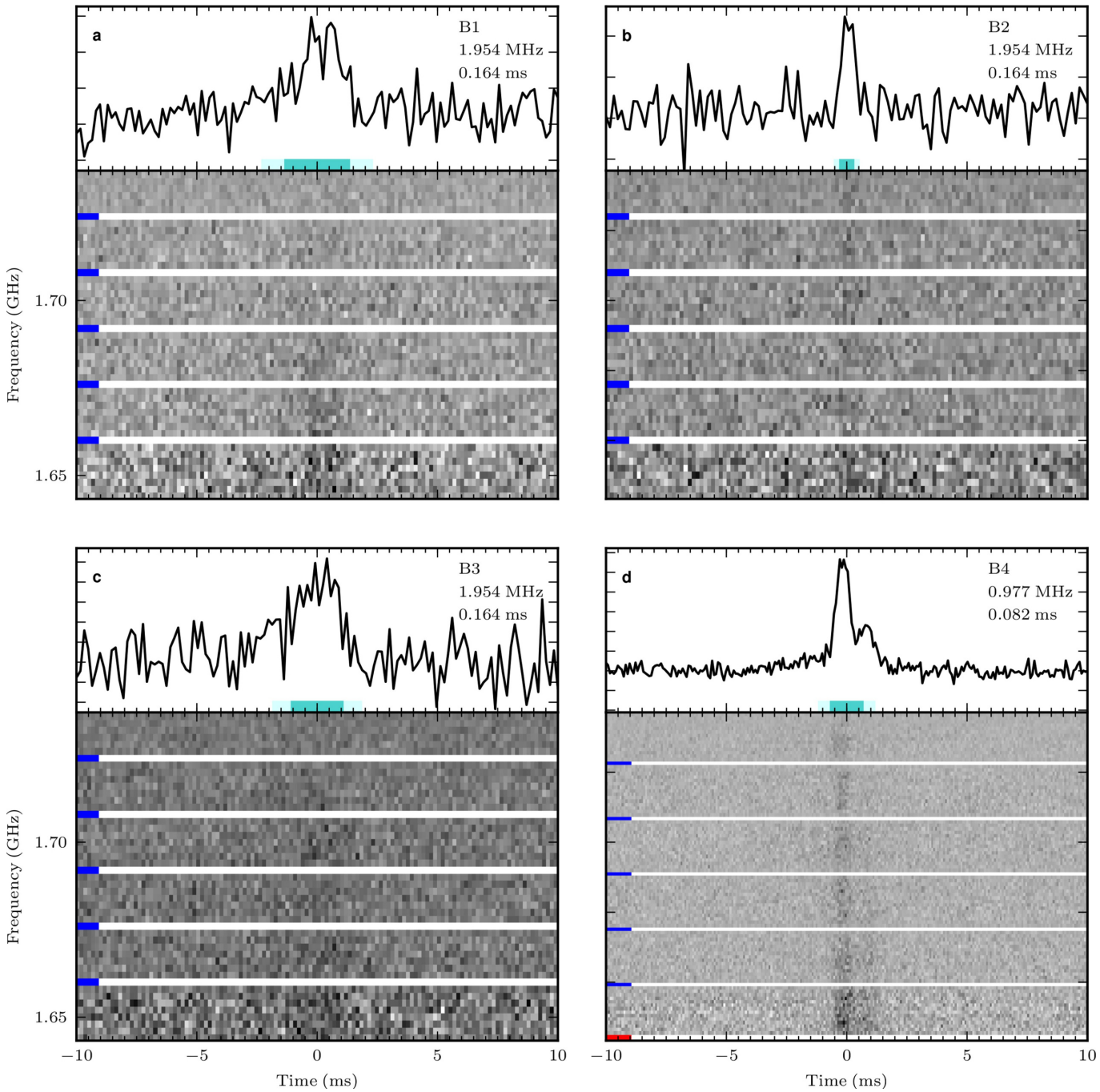
Competing interests The authors declare no competing interests.

Additional information

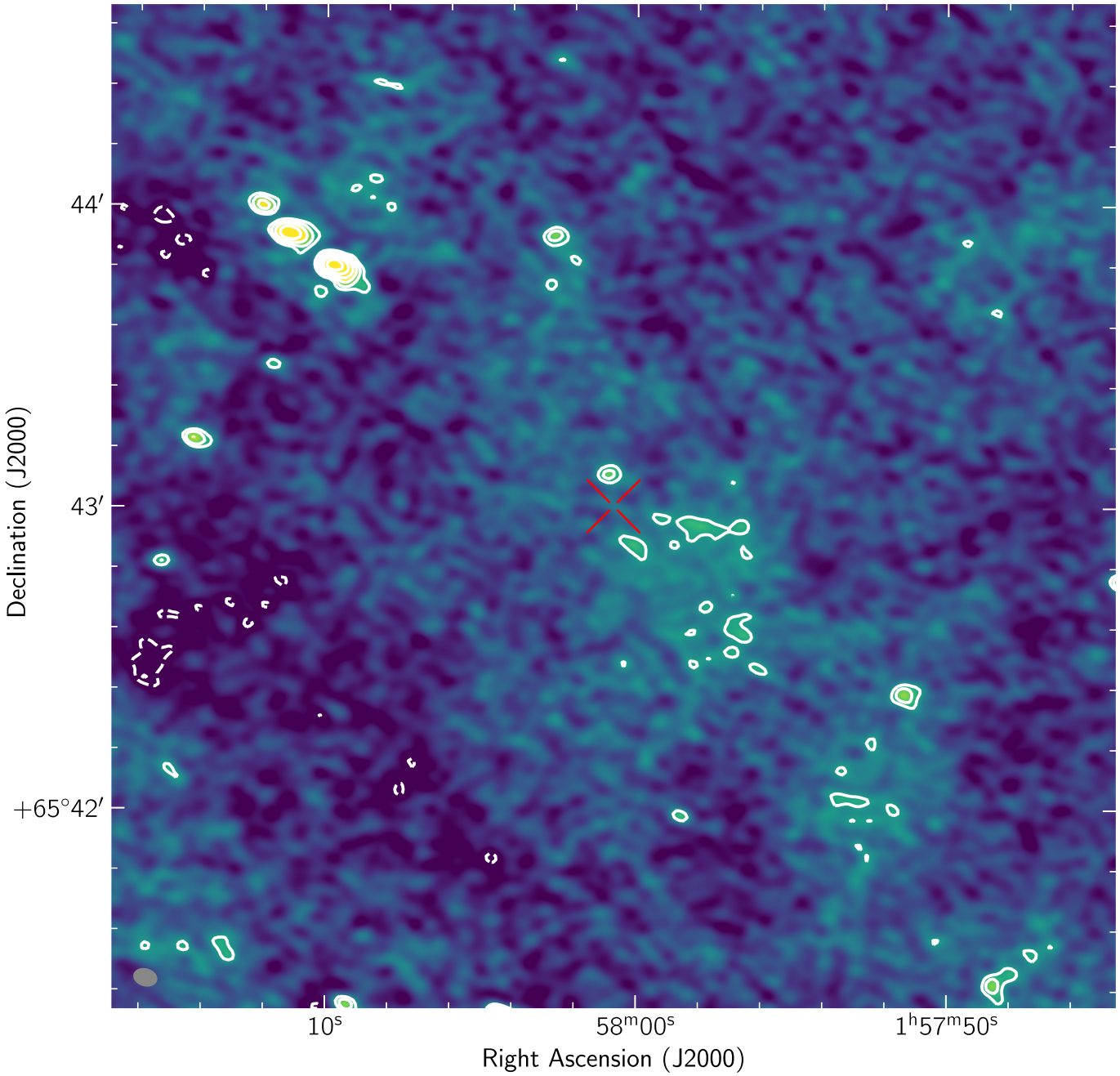
Correspondence and requests for materials should be addressed to J.W.T.H.

Peer review information *Nature* thanks Matthew Bailes and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

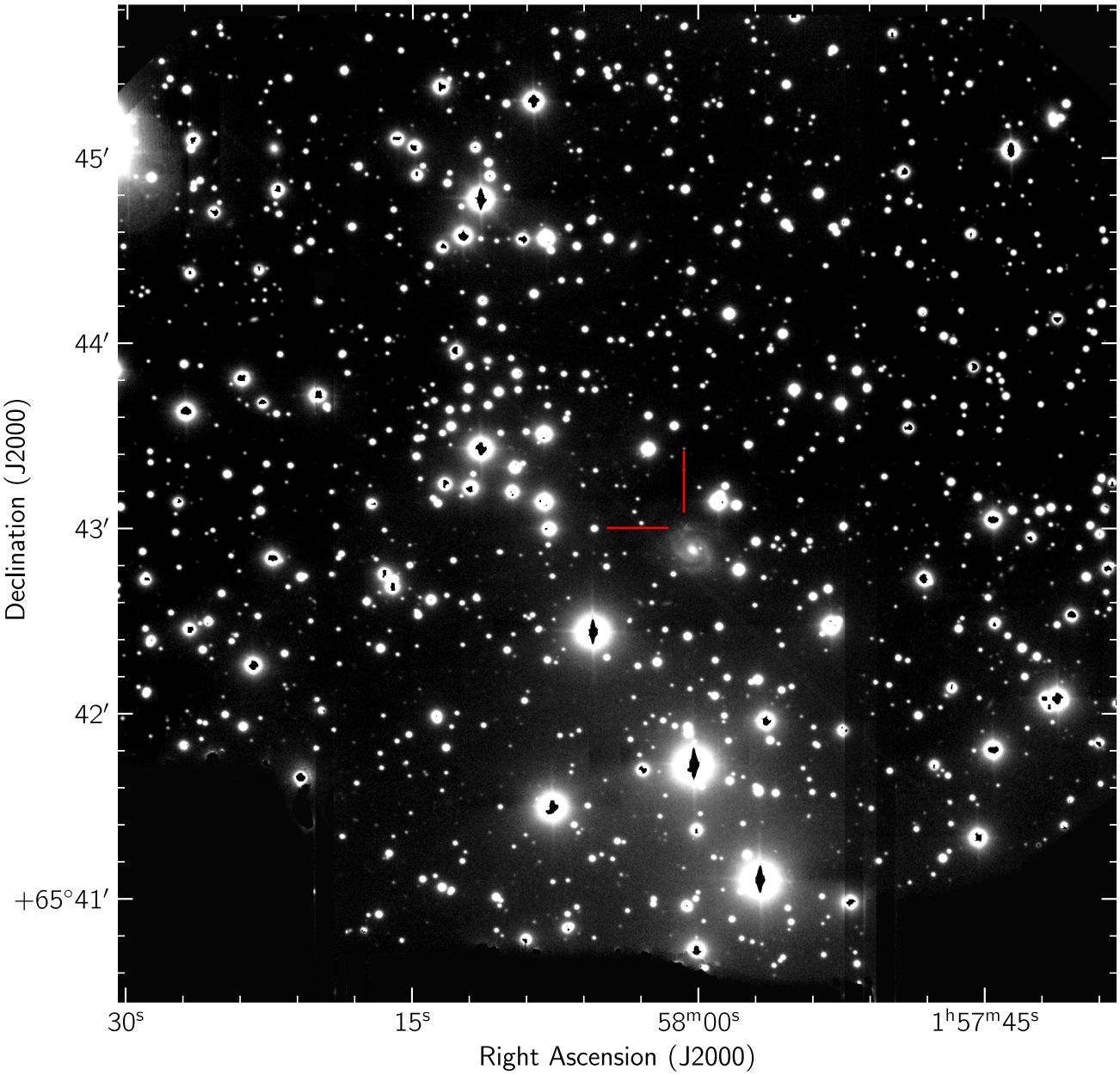
Reprints and permissions information is available at http://www.nature.com/reprints.



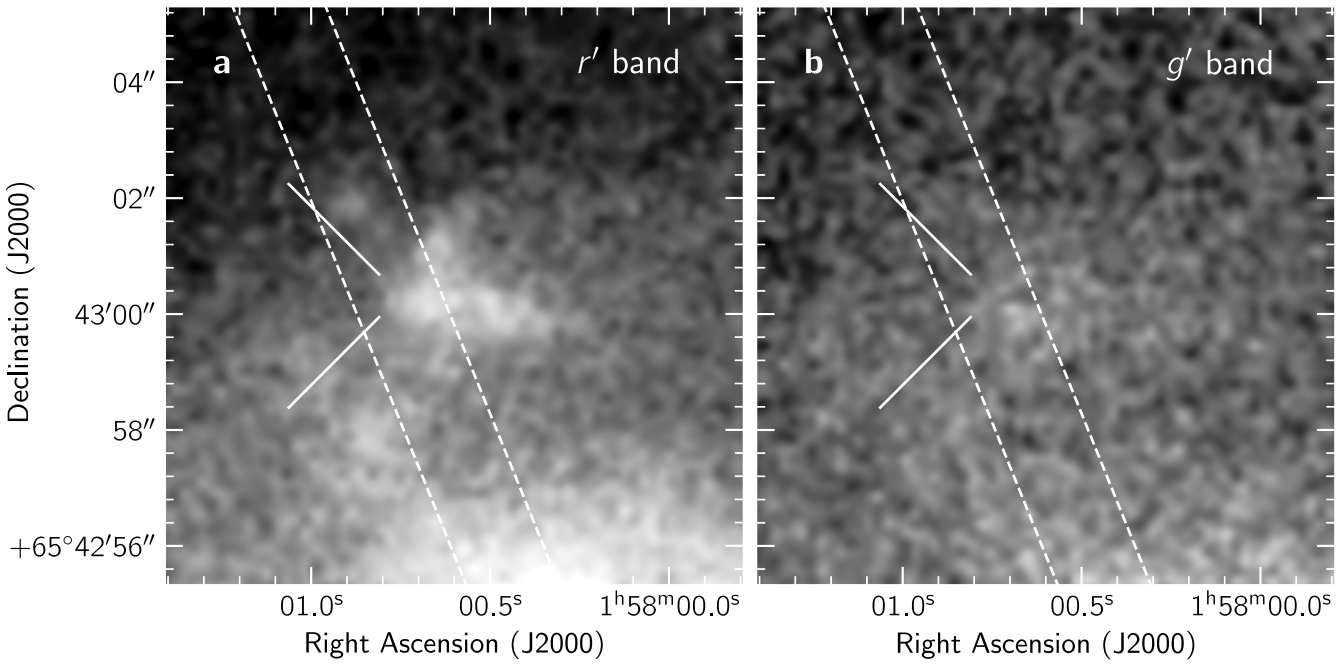
Extended Data Fig. 1 | Burst detections in Effelsberg PSRIX data. Band-averaged profiles and dynamic spectra of the four bursts, as detected in the PSRIX data (**a**, **b**, **c** and **d**). A 20-ms time window is shown surrounding the burst centre. Each burst was fitted with a Gaussian distribution to determine the FWHM duration, which is represented by the cyan bars. The lighter cyan encloses the 2σ region. The solid white lines are frequency channels that have been removed from the data due to either RFI or subband edges, indicated by the red and blue markers, respectively. For visual clarity, bursts B1, B2 and B3 (**a**, **b** and **c**) are downsampled in both time and frequency by a factor of two. The RFI excision was done before downsampling. The time and frequency resolution used for plotting is shown in the top right of each panel.



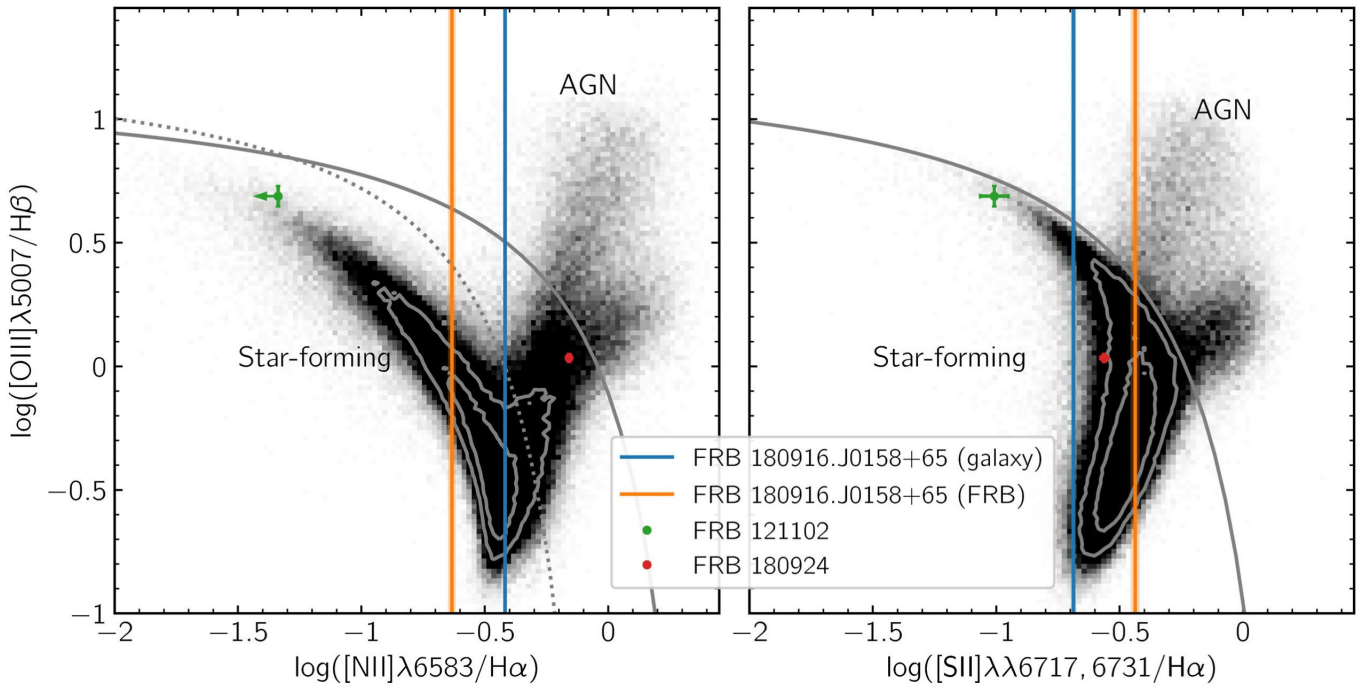
Extended Data Fig. 2 | VLA field image. Field of the continuum radio emission around FRB 180916.J0158+65 as seen by the VLA at 1.6 GHz with a bandwidth of 0.6 GHz. The position of FRB 180916.J0158+65 is marked by the red cross at the centre of the image. Contours start at the 3σ r.m.s. noise level of 18 μJy per beam and increase by factors of $\sqrt{2}$. The synthesized beam is represented by the grey ellipse in the bottom-left corner. Note that a faint source is detected at around 6 arcsec north of FRB 180916.J0158+65, but its separation is significant (>3σ confidence level) and we thus conclude that it is not associated with FRB 180916.J0158+65.



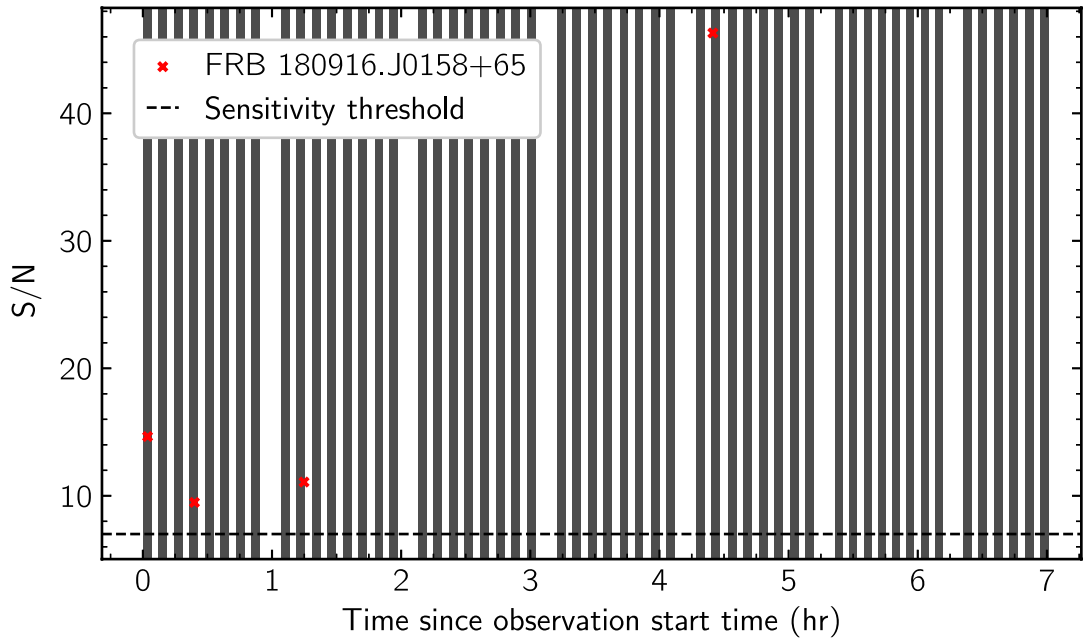
Extended Data Fig. 3 | Full field of view of the Gemini *r′* filter. The position of FRB 180916.J0158+65 is highlighted by the red cross. Note that the spiral galaxy associated with FRB 180916.J0158+65 is the only clearly visible galaxy in the field.



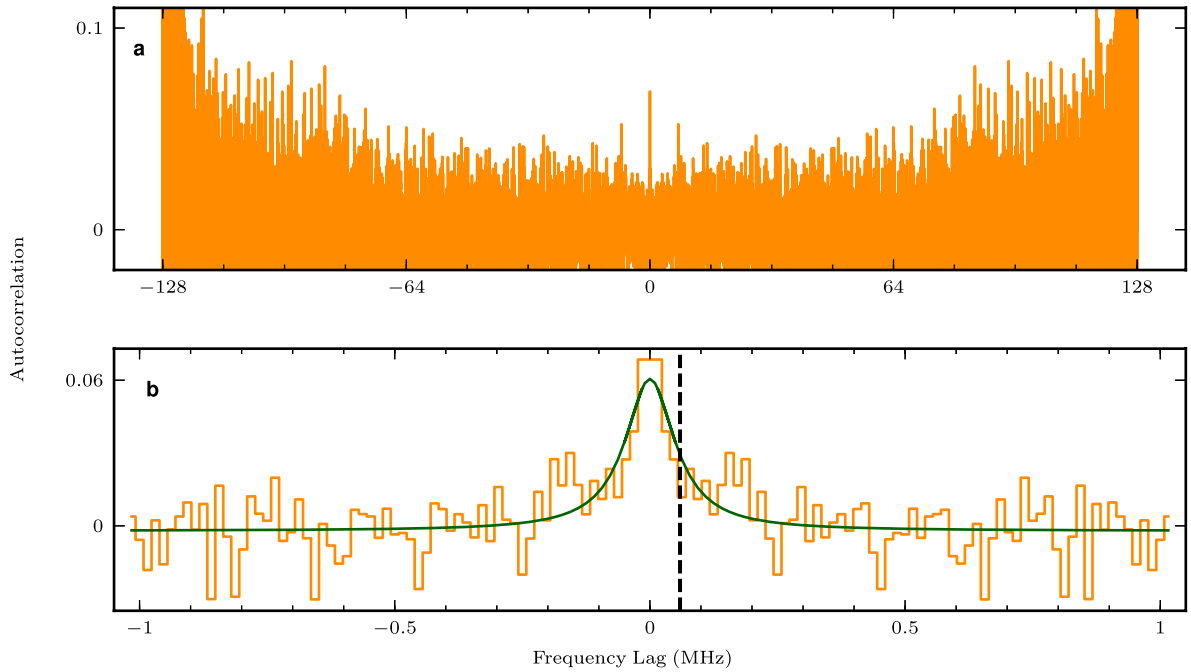
Extended Data Fig. 4 | Zoomed-in images at the position of FRB 180916.J0158+65. Gemini data at r' (**a**) and g' bands (**b**). The position of FRB 180916.J0158+65 is highlighted by the white cross. The uncertainty in its position is smaller than the resolution of these images. The dashed lines represent the orientation and placement of the 1.5-arcsec spectroscopic slit used to obtain the optical spectra. Note that the slit does not cover the full star-forming region but the region centred on FRB 180916.J0158+65, and that the whole region is strongly affected by extinction ($E(g-r) = 1.73 \pm 0.09$).



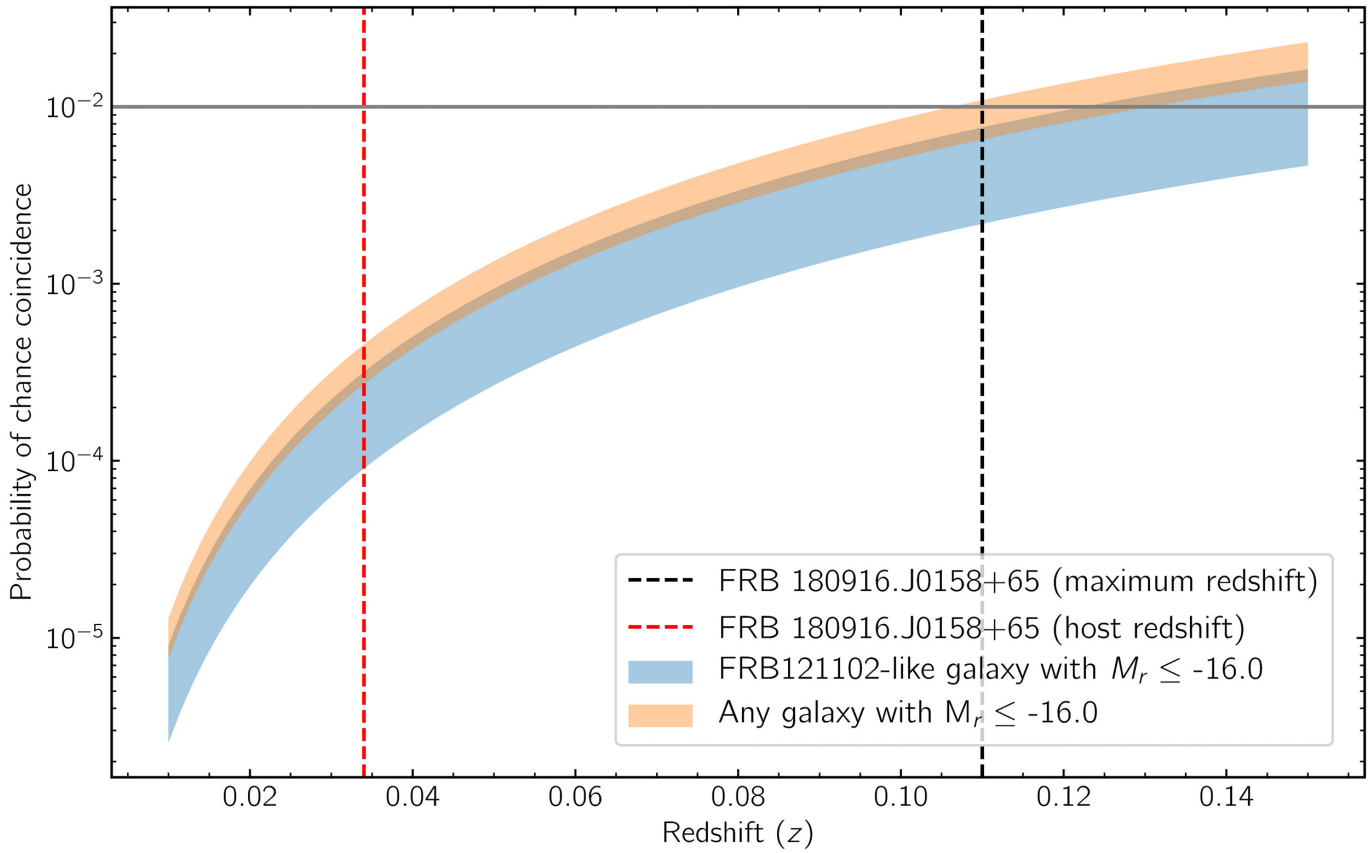
Extended Data Fig. 5 | Host galaxy source of ionization. Emission line flux ratios of [N II]/Hα (left) and [S II]/Hα (right) plotted against [O III]/Hβ. The greyscale distribution represents the SDSS DR12 sample[15] of 240,000 galaxies that display significant emission lines (>5σ), where the solid and dotted grey lines denote the demarcations between star-forming and AGN-dominated galaxies[73–75]. The host galaxies of FRB 121102 and FRB 180924 (shown by the green and red circles, respectively, where error bars show the 1σ uncertainty) are consistent with star-forming and AGN-dominated galaxies, respectively[11,17]. Though the Gemini-North spectrum of FRB 180916.J0158+65 does not cover the [O III] and Hβ lines, its [N II]/Hα and [S II]/Hα line ratios are broadly consistent with a star-formation-dominated galaxy (represented by the vertical lines and the 1σ region as linewidth).



Extended Data Fig. 6 | Burst brightness and arrival times. Burst *S*/*N* as a function of time during our 2019 June 19 observation of FRB 180916.J0158+65. The grey bars represent scans of the FRB 180916.J0158+65 field. The red crosses represent the four bursts (from left to right: B1, B2, B3, B4). The black dashed line indicates the detection threshold of our search in the pulsar-backend data (*S*/*N* = 7).



Extended Data Fig. 7 | Auto-correlation function and scintillation bandwidth of brightest burst, B4. a, The ACF of the spectrum of the bright, narrow burst component of burst B4. **b**, The ACF for lags between −1.016 MHz and +1.016 MHz. The zero-lag noise spike has been removed from the ACF. A Lorentzian fit is shown in green in **b**. The black vertical dashed line represents the scintillation bandwidth, defined as the half width at half maximum of the Lorentzian fit.



Extended Data Fig. 8 | Redshift-cumulated probability of chance alignment coincidence. Probability of a chance alignment between FRB 180916.J0158+65 and twice the median half-light radius of any galaxy with magnitude $M_B \leq -16$ (orange region) or with a dwarf galaxy like the host of FRB 121102 (blue region) as a function of redshift. The horizontal grey line represents the 1% probability threshold. At the redshift of the host galaxy, $z = 0.0337$ (vertical dashed red line), the chance coincidence probability is $P \ll 0.1\%$, and at the maximum possible redshift of about 0.11 derived from the observed DM the probability is <1% (vertical dashed black line).

Extended Data Table 1 | FRB 180916.J0158+65 burst properties as detected in both the Effelsberg PSRIX data and the Effelsberg auto-correlation data

		PSRIX data			Auto-correlations		
Burst	Peak time[a] (MJD)	S/N	DM (pc cm^{-3})	Width (ms)	S/N	DM (pc cm^{-3})	Width (ms)
B1	58653.0961366466	14.65	345	1.64	9.87	348.76	2.40
B2	58653.1112573504	9.48	349	0.33	9.61	348.76	0.32
B3	58653.1465969404	11.08	356	0.74	9.78	348.76	1.12
B4	58653.2785078914	46.30	350	0.49	65.42	348.76	0.06

S/N and width are determined using *single_pulse_search.py*, and are the exact values returned by this program for the best-fit boxcar function. The DM quoted for the PSRIX data are the DM corresponding to the optimal *S/N* detection. The DM for the Effelsberg auto-correlation data are 348.76 pc cm^{-3}, which is the *S/N* optimizing DM found for the brightest burst, B4, using the PSRCHIVE tool *pdmp*.

[a]Time of arrival (in Modified Julian Day, MJD) of the centre of the burst envelope at the Solar System barycentre after correcting to infinite frequency (that is, after removing the time delay from dispersion) using a DM of 348.76 pc cm^{-3}.

Extended Data Table 2 | Properties of the spectral emission lines

Line ID	Galaxy Core			FRB Location		
	Centroid (Å)	Width (Å)	Flux (10^{-16} erg cm^{-2} s^{-1})	Centroid (Å)	Width (Å)	Flux (10^{-16} erg cm^{-2} s^{-1})
O III 4958.93 Å	5122.7	5.5	0.38(4)	–	–	–
O III 5006.86 Å†	5155.6	5.6	4.76(4)	–	–	–
N II 6548.05 Å	6767.9	5.5	0.31(4)	–	–	–
Hα 6562.8 Å	6784.8	10.9	6.57(4)	6787.5	6.5	1.002(8)
N II 6583.45 Å	6806.9	10.3	2.51(4)	6809.0	5.1	0.233(8)
S II 6716.44 Å	6943.7	7.4*	0.77(4)	6946.2	6.4††	0.243(5)
S II 6730.81 Å	6953.8	7.4*	0.58(4)	6961.8	6.4††	0.123(5)

Properties of the most relevant spectral lines observed at the location of the core of the galaxy and at the position of FRB 180916.J0158+65. Note that the fluxes are not corrected for extinction. Numbers in parentheses indicate 1σ uncertainties in the least significant digits.

†The O III 5,006.86 Å line is detected at the FRB location but is corrupted by a cosmic-ray hit.

*,††A single value of linewidth was fitted for both lines.

Resonant microwave-mediated interactions between distant electron spins

https://doi.org/10.1038/s41586-019-1867-y

Received: 12 May 2019

Accepted: 30 September 2019

Published online: 25 December 2019

F. Borjans[1], X. G. Croot[1], X. Mi[1,2], M. J. Gullans[1] & J. R. Petta[1]*

Nonlocal qubit interactions are a hallmark of advanced quantum information technologies[1–5]. The ability to transfer quantum states and generate entanglement over distances much larger than qubit length scales greatly increases connectivity and is an important step towards maximal parallelism and the implementation of two-qubit gates on arbitrary pairs of qubits[6]. Qubit-coupling schemes based on cavity quantum electrodynamics[2,7,8] also offer the possibility of using high-quality-factor resonators as quantum memories[3,9]. Extending qubit interactions beyond the nearest neighbour is particularly beneficial for spin-based quantum computing architectures, which are limited by short-range exchange interactions[10]. Despite the rapidly maturing device technology for silicon spin qubits[11–16], experimental progress towards achieving long-range spin–spin coupling has so far been restricted to interactions between individual spins and microwave photons[17–20]. Here we demonstrate resonant microwave-mediated coupling between two electron spins that are physically separated by more than four millimetres. An enhanced vacuum Rabi splitting is observed when both spins are tuned into resonance with the cavity, indicating a coherent interaction between the two spins and a cavity photon. Our results imply that microwave-frequency photons may be used to generate long-range two-qubit gates between spatially separated spins.

The devices studied in this work consist of two double quantum dots (DQDs), denoted L-DQD (left DQD) and R-DQD (right DQD), that are fabricated on a Si/SiGe heterostructure and positioned at the left and right antinodes, respectively, of a half-wavelength Nb superconducting cavity (Fig. 1a). Device 1 is fabricated on a natural-silicon quantum well and has a cavity centre frequency of $f_c = 6.745$ GHz and a decay rate of $\kappa/(2\pi) = 1.98$ MHz. Device 2 uses an enriched ^{28}Si quantum well with a ^{29}Si residual concentration of 800 ppm. A single electron is isolated in each DQD[21] and interacts with the electric field of the cavity through the electric dipole interaction. Our device design uses a split-gate cavity coupler (labelled 'CP' in Fig. 1b) that is galvanically connected to the central pin of the superconducting cavity[22].

We first demonstrate the strong coupling of a spin trapped in each DQD to a cavity photon. In our device architecture, a large electric dipole coupling rate of $g_c/(2\pi) \approx 40$ MHz is combined with an artificial spin–orbit interaction generated by a micromagnet to achieve spin–photon coupling[17,23,24]. To probe the spin–photon coupling, the cavity transmission A/A_0 is plotted as a function of the cavity probe frequency f and B^{ext} in Fig. 1c. The Zeeman splitting increases with the total magnetic field $\mathbf{B}^{tot} = \mathbf{B}^{ext} + \mathbf{B}^{M}$, where $\mathbf{B}^{ext}$ is the externally applied field and $\mathbf{B}^{M}$ is the stray field of the micromagnet. Field tuning allows us to bring each spin into resonance with a cavity photon of energy hf_c, where h is Planck's constant. We observe coherent coupling between the spin trapped in the L-DQD and the cavity photon, as evidenced by the vacuum Rabi splitting when $B^{ext} = 109.1$ mT. Strong spin–photon coupling is achieved, with the spin–cavity coupling rate $g_{s,L}/(2\pi) = 10.7 \pm 0.1$ MHz (all error estimates in the manuscript correspond to one standard deviation) exceeding the cavity decay rate $\kappa/(2\pi) = 1.98$ MHz and the spin decoherence rate $\gamma_{s,L}/(2\pi) = 4.7$ MHz (see Supplementary Information). Likewise, we observe strong spin–photon coupling for the R (right) spin at $B^{ext} = 103.1$ mT, with $g_{s,R}/(2\pi) = 12.0 \pm 0.2$ MHz exceeding $\gamma_{s,R}/(2\pi) = 5.3$ MHz and $\kappa/(2\pi)$. Factoring in the susceptibility of the micromagnet, $\chi \approx 0.6$, the 6-mT difference in B^{ext} is equivalent to a 268-MHz difference in the spin resonance frequencies and precludes the observation of resonant coupling between the two spins[18].

To compensate for local differences in the magnetic field, the device was fabricated with the long axis of the Co micromagnets tilted by $\theta = \pm 15°$ relative to the interdot axis of the DQDs (Fig. 1b). Because of this intentional asymmetry, adjusting the angle φ of the in-plane magnetic field relative to the DQD axis provides an additional degree of freedom for simultaneous tuning of both spins into resonance with the cavity. Qualitatively, the high-permeability Co micromagnet concentrates the magnetic-field lines, leading to a maximum total field when B^{ext} is aligned with the long axis of the micromagnet[25]. We show that this technique is suitable for controlling the spin resonance frequencies in Fig. 1d, where we set the magnetic-field angle at $\varphi = 2.8°$. Here, the R spin resonance condition shifts up to $B^{ext} = 104.7$ mT and the L spin resonance condition shifts down to $B^{ext} = 107.4$ mT.

We now demonstrate control over the difference between the spin resonance frequencies by fixing the magnitude of the external magnetic field B^{ext} and varying the angle φ of B^{ext} in the plane of the sample. The expected spin resonance frequencies are plotted in Fig. 2a as a function of φ with $B^{ext} = 106.3$ mT (upper panel) and $B^{ext} = 110$ mT (lower panel), confirming that these two control parameters will bring the

[1]Department of Physics, Princeton University, Princeton, NJ, USA. [2]Present address: Google Inc., Santa Barbara, CA, USA. *e-mail: petta@princeton.edu

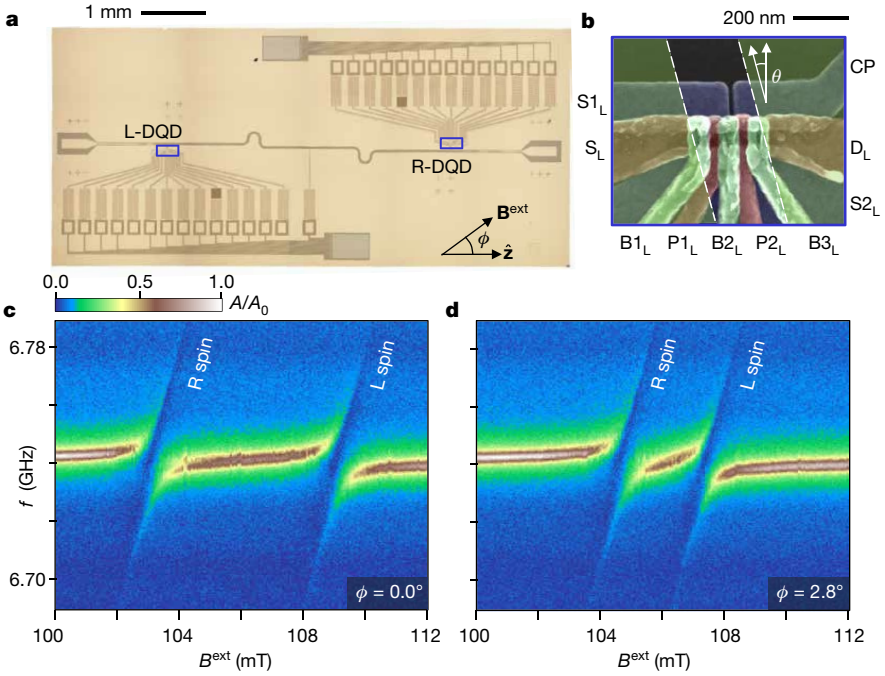


Fig. 1 | Cavity coupler for spins. a, Optical micrograph of the superconducting cavity containing two single-electron DQDs. The electron spin in each DQD is coupled to the cavity through a combination of electric-dipole and artificial spin–orbit interactions. $\hat{\mathbf{z}}$ indicates the direction of the DQD interdot axis. **b**, False-colour scanning electron microscope image of L-DQD. A double-well potential is formed beneath the plunger gates $P1_L$ and $P2_L$, and the barrier gate $B2_L$ is used to adjust interdot tunnel coupling. Spin–orbit coupling is induced by a Co micromagnet (dashed lines). CP, cavity coupler. $S1_L$, $S2_L$, S_L, D_L denote other gates used to define the L DQD. **c**, Cavity transmission A/A_0 as a function of B^{ext} and f. Vacuum Rabi splitting, a hallmark of strong coupling, is observed for each spin. For a field angle of $\varphi = 0°$ the R spin and L spin resonance conditions are separated by $\Delta B^{ext} = 6$ mT. **d**, Increasing the field angle to $\varphi = 2.8°$ reduces the field separation drastically.

two spins into resonance with the cavity and with each other. On the basis of microwave spectroscopy measurements of the spins, we expect resonance to occur around $\varphi = 6°$ and $B^{ext} = 106.3$ mT. Figure 2b maps out the field angle dependence of the spin resonance frequencies over the range $\varphi = 3°–8°$ with $B^{ext} = 106.3$ mT. The resonance frequency of the R spin monotonically moves to lower f values as φ is increased, whereas the L spin shows the opposite dependence. With $\varphi = 5.6°$ both spins are tuned into resonance with the cavity. These results are well captured by the theoretical prediction of transmission through the cavity shown in Fig. 2c.

The spectrum of the Jaynes–Cummings model for a single spin and a single photon in the cavity is shown in Fig. 3a. With the L spin tuned into resonance with the cavity, the spin and cavity photon hybridize, leading to a vacuum Rabi splitting of magnitude $2g_{s,L}$ in the cavity transmission[7]. By contrast, when both spins are simultaneously tuned into resonance with the cavity, the excited-state spectrum of the

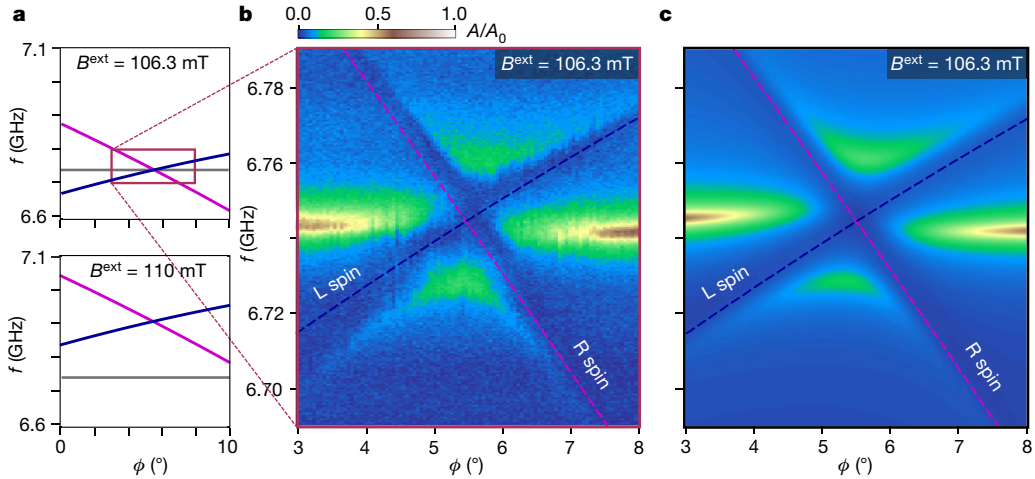


Fig. 2 | Tuning two spatially separated spins into resonance. a, Expected spin resonance frequencies as a function of φ for $B^{ext} = 106.3$ mT (top panel) and $B^{ext} = 110$ mT (bottom panel). B^{ext} allows control over both spin frequencies with respect to the cavity, whereas φ allows control over the spin frequencies with respect to each other. The frequency of the L (R) spin is plotted in blue (purple). **b**, A/A_0 as a function of f and φ, demonstrating simultaneous tuning of both spins into resonance with the cavity at $\varphi = 5.6°$ and $B^{ext} = 106.3$ mT. Dashed lines indicate L spin and R spin transition frequencies. **c**, Theoretical prediction for A/A_0.

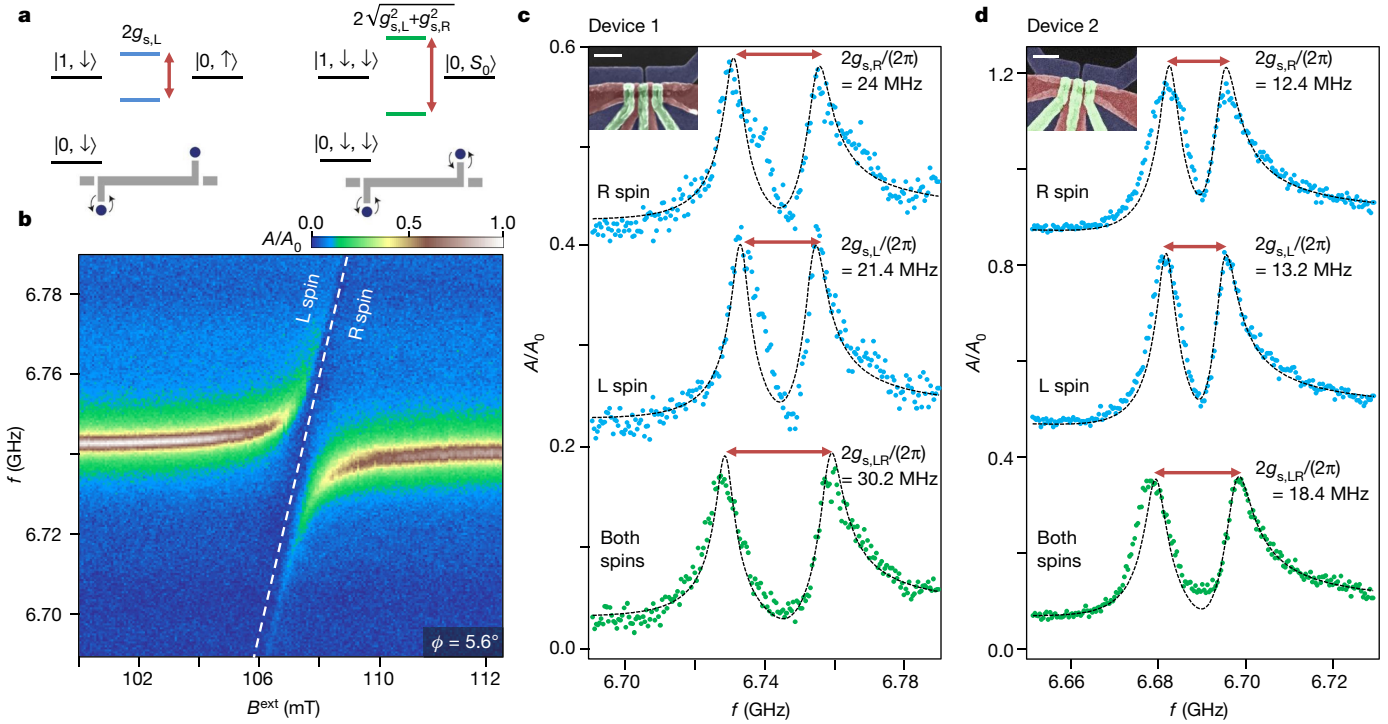


Fig. 3 | Resonant coupling of the two spins via a cavity photon. a, Tuning the L spin into resonance with the cavity results in vacuum Rabi splitting with magnitude $2g_{s,L}$. A vacuum Rabi splitting of magnitude $2g_{s,LR} = 2\sqrt{g_{s,L}^2 + g_{s,R}^2}$ is expected when both spins are tuned into resonance with the cavity. **b**, A/A_0 as a function of f and B^{ext} with $\varphi = 5.6°$, indicating an enhanced vacuum Rabi splitting when the L spin and R spin are tuned into resonance with the cavity. **c**, A/A_0 as a function of f for the R spin in resonance (upper curve), the L spin in resonance (middle curve) and both spins in resonance (bottom curve). The enhancement of the vacuum Rabi splitting with both spins in resonance with the cavity is indicative of coupling between the two spins via the cavity mode. **d**, Cavity-assisted spin coupling is also observed in a second device with a different gate pattern. Dashed lines in **c**, **d** are fits to a master-equation simulation. Insets in **c**, **d** are scanning electron microscope images of the devices, with scale bars of 200 nm.

Jaynes–Cummings model splits into a sub-radiant state and two bright states[2]. For N identical spins, the Jaynes–Cummings model predicts a collective $\sqrt{N}$ enhancement of the coupling rate[7]. In our device geometry, the sub-radiant state is the spin triplet $|0, T_0\rangle = (1/\sqrt{2})(|0,\uparrow,\downarrow\rangle + |0,\downarrow,\uparrow\rangle)$ because the DQDs are located at opposite ends of the cavity, where the electric fields are 180° out of phase. Here, the spin states of the L/R spin are quantized along the axis of their local total magnetic field. The bright states are hybridizations between the singlet state $|0, S_0\rangle = (1/\sqrt{2})(|0,\uparrow,\downarrow\rangle - |0,\downarrow,\uparrow\rangle)$ and the state with a single photon $|1,\downarrow,\downarrow\rangle$, which are separated in energy by twice the collectively enhanced vacuum Rabi coupling, $2g_{s,LR} = 2\sqrt{g_{s,L}^2 + g_{s,R}^2}$. We now search for evidence of cavity-mediated single-spin coupling.

Figure 3b shows A/A_0 as a function of f and B^{ext} with $\varphi = 5.6°$, where both spins are in resonance with the cavity. As B^{ext} is increased, both spins are simultaneously tuned into resonance with the cavity at $B^{ext} = 106.3$ mT and we observe an enhancement in the vacuum Rabi splitting relative to the datasets shown in Fig. 1c, d. The vacuum Rabi splitting is quantitatively analysed for device 1 in Fig. 3c and for device 2 in Fig. 3d. These devices have slight differences in gate geometry, and the micromagnets in device 2 are canted at angles of $\theta = \pm 10°$. In device 1, we extract $g_{s,L/R}$ by measuring A/A_0 at $\varphi = 5.6°$ by decoupling the R/L spin from the cavity. This is achieved by increasing the charge level detuning ($\varepsilon_R/\varepsilon_L$) in the respective DQD to localize the electron in a single quantum dot[18]. In device 2, we extract $g_{s,L/R}$ by measuring A/A_0 at an off-resonance angle, at the now separated resonant fields (see Supplementary Fig. 1). For device 1, we observe a vacuum Rabi splitting of $2g_{s,L}/(2\pi) = 21.4 \pm 0.2$ MHz when the L spin is in resonance with the cavity, $2g_{s,R}/(2\pi) = 24.0 \pm 0.4$ MHz when the R spin is in resonance with the cavity, and $2g_{s,LR}/(2\pi) = 30.2 \pm 0.2$ MHz when both spins are in resonance with the cavity. Device 2 shows similar behaviour and again exhibits an enhanced vacuum Rabi splitting of $2g_{s,LR}/(2\pi) = 18.4 \pm 0.4$ MHz when both spins are tuned into resonance with the cavity, compared to the individual splittings $2g_{s,L}/(2\pi) = 13.2 \pm 0.2$ MHz and $2g_{s,R}/(2\pi) = 12.4 \pm 0.2$ MHz. Combined, these two datasets give strong evidence for microwave-assisted interactions between two spins across a length scale of 4 mm, which is many orders of magnitude larger than what can be achieved using direct wavefunction overlap. Moreover, these measurements show that both the field angle and DQD-level detuning can be used to modulate the strength of the interactions between the two spins.

The data in Fig. 3c, d are fitted using a master-equation description of the spin–cavity system (see Supplementary Information). We independently measure the linewidths $\gamma_{s,L/R}$ using microwave spectroscopy. The cavity linewidth κ and the complex Fano factor q are obtained by fitting the bare cavity response with the spins detuned from resonance. The spin–photon coupling rates $g_{s,L/R}$ for each device are obtained by fitting the data obtained with the spins individually tuned into resonance with the cavity, as shown in Fig. 3c, d. From the Jaynes–Cummings model we expect the bright states to split with the enhanced collective coupling rate $2g_{s,LR} = 2\sqrt{g_{s,L}^2 + g_{s,R}^2}$. The extracted splittings agree with the theoretical predictions of $2g_{s,LR}/(2\pi) = 32.1$ MHz for device 1 and $2g_{s,LR}/(2\pi) = 18.1$ MHz for device 2 to within 6%.

The observation of enhanced vacuum Rabi splitting when the separated spins are simultaneously in resonance with the cavity is evidence of a long-range interaction between the two spins and a cavity photon. The nonlocal interaction of two spins marks an important milestone for all-to-all qubit connectivity and scalability in silicon-based quantum circuits. With further improvements in cavity quality factors and spin–photon coupling rates[18], time-domain demonstrations of cavity-assisted spin–spin coupling should be within experimental reach. The spin–photon coupling rate could be considerably increased by using high-impedance superconducting cavities[19,20], potentially enabling

spin–spin coupling in the dispersive regime, where the spins are detuned from the cavity and the photonic degree of freedom can be effectively eliminated[2]. High-fidelity long-range spin–spin coupling in the dispersive regime would allow the implementation of modular qubit architectures in silicon, where nearest-neighbour coupled registers of spin qubits[22,26] can be interfaced with other sparsely distributed registers via microwave photons.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1867-y.

1. Duan, L.-M., Lukin, M. D., Cirac,m J. I. & Zoller, P. Long-distance quantum communication with atomic ensembles and linear optics. *Nature* **414**, 413–418 (2001).
2. Majer, J. et al. Coupling superconducting qubits via a cavity bus. *Nature* **449**, 443–447 (2007).
3. Sillanpää, M. A., Park, J. I. & Simmonds, R. W. Coherent quantum state storage and transfer between two phase qubits via a resonant cavity. *Nature* **449**, 438–442 (2007).
4. Monroe, C. & Kim, J. Scaling the ion trap quantum processor. *Science* **339**, 1164–1169 (2013).
5. Axline, C. J. et al. On-demand quantum state transfer and entanglement between remote microwave cavity memories. *Nat. Phys.* **14**, 705–710 (2018).
6. Preskill, J. Reliable quantum computers. *Proc. R. Soc. Lond. A* **454**, 385–410 (1998).
7. Wallraff, A. et al. Strong coupling of a single photon to a superconducting qubit using circuit quantum electrodynamics. *Nature* **431**, 162–167 (2004).
8. Trif, M., Golovach, V. N. & Loss, D. Spin dynamics in InAs nanowire quantum dots coupled to a transmission line. *Phys. Rev. B* **77**, 045434 (2008).
9. Pfaff, W. et al. Controlled release of multiphoton quantum states from a microwave cavity memory. *Nat. Phys.* **13**, 882–887 (2017).
10. Petta, J. R. et al. Coherent manipulation of coupled electron spins in semiconductor quantum dots. *Science* **309**, 2180–2184 (2005).
11. Yoneda, J. et al. A quantum-dot spin qubit with coherence limited by charge noise and fidelity higher than 99.9%. *Nat. Nanotechnol.* **13**, 102–106 (2018).
12. Veldhorst, M. et al. A two-qubit logic gate in silicon. *Nature* **526**, 410–414 (2015).
13. Zajac, D. M. et al. Resonantly driven CNOT gate for electron spins. *Science* **359**, 439–442 (2018).
14. Watson, T. F. et al. A programmable two-qubit quantum processor in silicon. *Nature* **555**, 633–637 (2018).
15. Fujita, T., Baart, T. A., Reichl, C., Wegscheider, W. & Vandersypen, L. M. K. Coherent shuttle of electron-spin states. *npj Quant. Inf.* **3**, 22 (2017).
16. Mills, A. R. et al. Shuttling a single charge across a one-dimensional array of silicon quantum dots. *Nat. Commun.* **10**, 1063 (2019).
17. Viennot, J. J., Dartiailh, M. C., Cottet, A. & Kontos, T. Coherent coupling of a single spin to microwave cavity photons. *Science* **349**, 408–411 (2015).
18. Mi, X. et al. A coherent spin–photon interface in silicon. *Nature* **555**, 599–603 (2018).
19. Samkharadze, N. et al. Strong spin-photon coupling in silicon. *Science* **359**, 1123–1127 (2018).
20. Landig, A. J. et al. Coherent spin–photon coupling using a resonant exchange qubit. *Nature* **560**, 179–184 (2018).
21. Loss, D. & DiVincenzo, D. P. Quantum computation with quantum dots. *Phys. Rev. A* **57**, 120–126 (1998).
22. Zajac, D. M., Hazard, T. M., Mi, X., Nielsen, E. & Petta, J. R. Scalable gate architecture for a one-dimensional array of semiconductor spin qubits. *Phys. Rev. Appl.* **6**, 054013 (2016).
23. Cottet, A. & Kontos, T. Spin quantum bit with ferromagnetic contacts for circuit QED. *Phys. Rev. Lett.* **105**, 160502 (2010).
24. Benito, M., Mi, X., Taylor, J. M., Petta, J. R. & Burkard, G. Input-output theory for spin-photon coupling in Si double quantum dots. *Phys. Rev. B* **96**, 235434 (2017).
25. Astner, T. et al. Coherent coupling of remote spin ensembles via a cavity bus. *Phys. Rev. Lett.* **118**, 140502 (2017).
26. Vandersypen, L. M. K. et al. Interfacing spin qubits in quantum dots and donors—hot, dense, and coherent. *npj Quant. Inf.* **3**, 34 (2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Data availability

The data that support the findings of this study are available from the corresponding author on reasonable request.

Acknowledgements This work was funded by Army Research Office grant W911NF-15-1-0149 and the Gordon and Betty Moore Foundation's EPiQS Initiative through grant GBMF4535. The devices were fabricated in the Princeton University Quantum Device Nanofabrication Laboratory. We thank L. Edge, J. Kerckhoff, T. Ladd and E. Pritchett of HRL Laboratories, LLC for providing the ^{28}Si heterostructure used in these experiments, for device simulation support and for technical comments on the manuscript. We acknowledge discussions with M. Benito and G. Burkard.

Author contributions F.B. and X.G.C. carried out the measurements with input from X.M. and J.R.P.; F.B. and X.M. fabricated the device. M.J.G. provided theory support. F.B., X.G.C., M.J.G. and J.R.P. wrote the manuscript. All authors discussed the results and commented on the manuscript.

Competing interests X.M., J.R.P. and Princeton University have filed a non-provisional patent application related to spin–photon transduction (US patent application number 16534431).

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1867-y.

Correspondence and requests for materials should be addressed to J.R.P.

Peer review information *Nature* thanks Hendrik Bluhm and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

Reprints and permissions information is available at http://www.nature.com/reprints.

Article

Synthesis and properties of free-standing monolayer amorphous carbon

https://doi.org/10.1038/s41586-019-1871-2

Received: 10 November 2018

Accepted: 26 September 2019

Published online: 8 January 2020

Chee-Tat Toh[1,2,10], Hongji Zhang[3,10], Junhao Lin[4,5], Alexander S. Mayorov[2], Yun-Peng Wang[6,7], Carlo M. Orofeo[1], Darim Badur Ferry[1], Henrik Andersen[2], Nurbek Kakenov[2], Zenglong Guo[5], Irfan Haider Abidi[2], Hunter Sims[6], Kazu Suenaga[4,8], Sokrates T. Pantelides[6,9] & Barbaros Özyilmaz[1,2,3]*

Bulk amorphous materials have been studied extensively and are widely used, yet their atomic arrangement remains an open issue. Although they are generally believed to be Zachariasen continuous random networks[1], recent experimental evidence favours the competing crystallite model in the case of amorphous silicon[2–4]. In two-dimensional materials, however, the corresponding questions remain unanswered. Here we report the synthesis, by laser-assisted chemical vapour deposition[5], of centimetre-scale, free-standing, continuous and stable monolayer amorphous carbon, topologically distinct from disordered graphene. Unlike in bulk materials, the structure of monolayer amorphous carbon can be determined by atomic-resolution imaging. Extensive characterization by Raman and X-ray spectroscopy and transmission electron microscopy reveals the complete absence of long-range periodicity and a threefold-coordinated structure with a wide distribution of bond lengths, bond angles, and five-, six-, seven- and eight-member rings. The ring distribution is not a Zachariasen continuous random network, but resembles the competing (nano)crystallite model[6]. We construct a corresponding model that enables density-functional-theory calculations of the properties of monolayer amorphous carbon, in accordance with observations. Direct measurements confirm that it is insulating, with resistivity values similar to those of boron nitride grown by chemical vapour deposition. Free-standing monolayer amorphous carbon is surprisingly stable and deforms to a high breaking strength, without crack propagation from the point of fracture. The excellent physical properties of this stable, free-standing monolayer amorphous carbon could prove useful for permeation and diffusion barriers in applications such as magnetic recording devices and flexible electronics.

Amorphous materials are used in a wide range of applications, but their atomic-scale structure and its effect on properties are far more complex than those of crystalline analogues. In a classic 1932 paper, Zachariasen[1] invoked free-energy arguments to propose that amorphous materials comprise the same bonding units as their crystalline analogues, but these units form continuous random networks (Z-CRNs) instead of periodic structures. A competing model, namely, the existence of crystallites embedded in an otherwise CRN environment, is even older[7] (the crystallite model of amorphous solids is distinct from nanocrystalline materials, as the latter comprise nanoscale grains separated by grain boundaries). In the past few decades, the Z-CRN model has gained wide acceptance, especially for amorphous Si and SiO_2 (a-Si and a-SiO_2), which are viewed as prototypes. The primary experimental evidence was provided by radial distribution functions, which could be reproduced accurately by model Z-CRNs. More recently, however, experiments that probe beyond pair correlations have found extensive concentrations of nanocrystallites of order 1–2 nm, whereas the corresponding models fit the experimental radial distribution functions just as well as Z-CRN models[2–4]. The inability to image the atomic-scale structure of bulk amorphous materials directly by microscopy means that the debate over CRNs versus crystallites has remained unresolved[8].

For two-dimensional (2D) atomically thin materials, the nature of the amorphous state can in principle be resolved by direct atomic-resolution imaging. Monolayer amorphous carbon (MAC) can be viewed as a prototype amorphous 2D material, an analogue of monolayer crystalline carbon (graphene). Atomic-resolution transmission electron

[1]Department of Physics, National University of Singapore, Singapore, Singapore. [2]Centre for Advanced 2D Materials, National University of Singapore, Singapore, Singapore. [3]Department of Materials Science and Engineering, National University of Singapore, Singapore, Singapore. [4]National Institute of Advanced Industrial Science and Technology, Tsukuba, Japan. [5]Department of Physics, Southern University of Science and Technology, Shenzhen, China. [6]Department of Physics and Astronomy, Vanderbilt University, Nashville, TN, USA. [7]Hunan Key Laboratory for Super-Microstructure and Ultrafast Process, School of Physics and Electronics, Central South University, Changsha, China. [8]Department of Mechanical Engineering, University of Tokyo, Tokyo, Japan. [9]Department of Electrical Engineering and Computer Science, Vanderbilt University, Nashville, TN, USA. [10]These authors contributed equally: Chee-Tat Toh, Hongji Zhang. *e-mail: phyob@nus.edu.sg

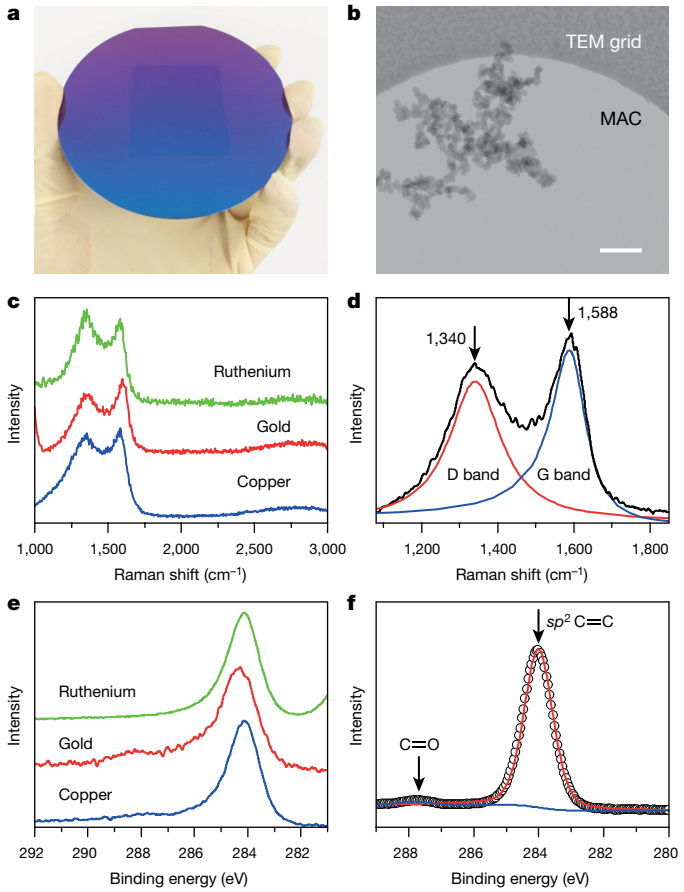


Fig. 1 | Morphology of MAC. a, MAC, 5 × 5 cm^2, transferred onto a SiO_2/Si wafer. **b**, Scanning electron microscope (SEM) image of MAC suspended on a TEM grid has uniform contrast throughout. Silicon nanoparticles have been intentionally suspended on the MAC to emphasize the presence of the film. Scale bar, 200 nm. **c**–**f**, Characterization of MAC grown on different substrates. **c**, Raman spectra for MAC grown on copper and gold substrates were measured after transferring to SiO_2. The Raman spectrum for MAC on ruthenium was measured directly on the growth substrate. **d**, Raman spectra for growth on copper, showing the D and G bands with the fitted curve and an I_D/I_G ratio of 0.82. **e**, C 1*s* XPS spectra measured directly on different growth substrates. **f**, High-resolution C 1*s* XPS spectra on copper, with the fitted curve showing a single carbon sp^2 peak at 284.0 eV.

microscopy (TEM) has been used to image freestanding graphene monolayers irradiated with TEM electron beams to induce disorder[9,10]. Such monolayers, however, are limited in size, generally inhomogeneous because of carbon loss and, when crystallites are still present, they are not necessarily randomly oriented, as they would be in a truly amorphous material, and can be eliminated by further irradiation. Similarly, radiation of aromatic self-assembled monolayers on substrates produces nanocrystalline graphene with only small amorphous patches[11].

Direct synthesis of MAC samples remains a challenge. Recently, the synthesis of amorphous carbon monolayers on Ge substrates using conventional chemical vapour deposition (CVD) was reported[12]. These samples were grown at high temperatures (>900 °C), resulting in amorphous regions (about 300 nm) that are embedded in a graphene matrix. The reported TEM images lack atomic resolution to establish the structure unambiguously. Instead, the authors concluded in favour of a Z-CRN structure on the basis of the halos in diffraction patterns derived from fast Fourier transform patterns of TEM images of regions 2.5 × 2.5 nm^2. However, many of the halos exhibit residual six-dot motifs—that is, the existence of crystallites 1–2 nm across cannot be ruled out.

Here we report the synthesis of MAC by a laser-assisted CVD growth process. The self-limiting process leads to a uniform monolayer of several square centimetres in less than 1 minute at substrate temperatures as low as 200 °C. Extensive atomic-resolution TEM characterization shows unambiguously both the complete absence of long-range periodicity and a structure that is consistent with the crystallite model: that is, nanometre-sized, randomly oriented and strained crystallites comprising only six-member rings embedded in a Z-CRN environment. For simplicity, here we discuss representative data on MAC films grown on a copper foil at 250 °C unless otherwise stated. Samples are easily transferred via wet etching after growth. Unlike other CVD-grown 2D materials, MAC is freestanding even on liquid surfaces without the need of a support polymer (Supplementary Video). Figure 1a, b shows that the transferred sample (on a SiO_2/Si wafer and a TEM grid, respectively) is homogeneous and continuous, free from multilayer regions or the wrinkles that are typically observed with the transfer of 2D layers. In addition, MAC is stable and freestanding even when stored under ambient conditions for at least 1 year.

The MAC samples were first characterized by Raman spectroscopy and X-ray photoemission spectroscopy (XPS) measurements. All samples have nearly identical spectra independent of substrate (Fig. 1c, d). The Raman 2D band at approximately 2,680 cm^{-1}, which is pronounced in crystalline graphene, is negligible in the MAC, strongly suggesting the lack of any long-range order[13]. Furthermore, the Raman intensity ratio $I_D/I_G = 0.82$ (Fig. 1e) indicates an average defect distance[14] of <1 nm. Note that Raman mapping of I_D/I_G shows that MAC is uniform over areas of 50 × 50 μm^2 (Extended Data Fig. 1). Furthermore, C 1*s* XPS in Fig. 1f shows that the bonds in MAC are mainly carbon sp^2 (Supplementary Note 1).

Monochromated, aberration-corrected, high-resolution transmission electron microscopy (HRTEM) was used to image the exact arrangement of carbon atoms in MAC. A large-area HRTEM image reveals a connected but distorted structure of five-, six-, seven- and eight-member rings (Fig. 2a). The Fourier transform in the inset of Fig. 2a shows a broad, continuous halo. Figure 2b shows a zoomed-in area of 5 × 5 nm^2 in which we can clearly see heavily distorted crystallites (green regions) about 1 nm across, embedded in a CRN background that comprises five-, six-, seven- and eight-member rings. As the yellow arrows indicate, the crystallites are oriented randomly. This structure of MAC is consistent over the entire sample (Supplementary Note 2). Thus, MAC is topologically distinct from disordered graphene.

We also evaluated the pair correlation function of neighbouring carbon atoms in MAC samples (Fig. 2e) and a graphene reference (Supplementary Note 3), a crucial property in deciding whether a material is amorphous. Graphene peaks disappear or are highly broadened in the MAC pair correlation function after the second nearest neighbours, confirming the loss of long-range periodic order. The bond lengths and angles for graphene are centred at 1.4 Å and 120°, respectively, with small variation arising from image aberration and algorithmic error. In contrast, MAC has much broader variation, 0.9–1.8 Å and 90–150° for the in-plane projections of bond lengths and bond angles, respectively (Fig. 2f, g). This feature is surprising because the theoretical breaking strain at 25–30% for crystalline graphene occurs with deformations[15] at just under 1.6 Å and 135°, so one would expect freestanding MAC to be unstable and subject to collapse. On the other hand, such a wide spread in the bond angle distribution is expected for a random 2D amorphous network, similar to the O–Si–O bonds modelled in a silica bilayer on ruthenium[16]. Moreover, each carbon atom is three-fold coordinated, consistent with the sp^2 bonding concluded from the XPS data.

Furthermore, the strained crystallites are fundamentally different from nanocrystalline graphene domains because they are separated not by atomically sharp grain boundaries, but by CRN regions that are at least three carbon atoms wide (Fig. 2h, i). Selective-area electron diffraction (SAED) patterns show a characteristic diffuse halo (Fig. 2j) that confirms the amorphous nature of MAC, in contrast to the sharp first- and second-order diffraction rings for nanocrystalline graphene

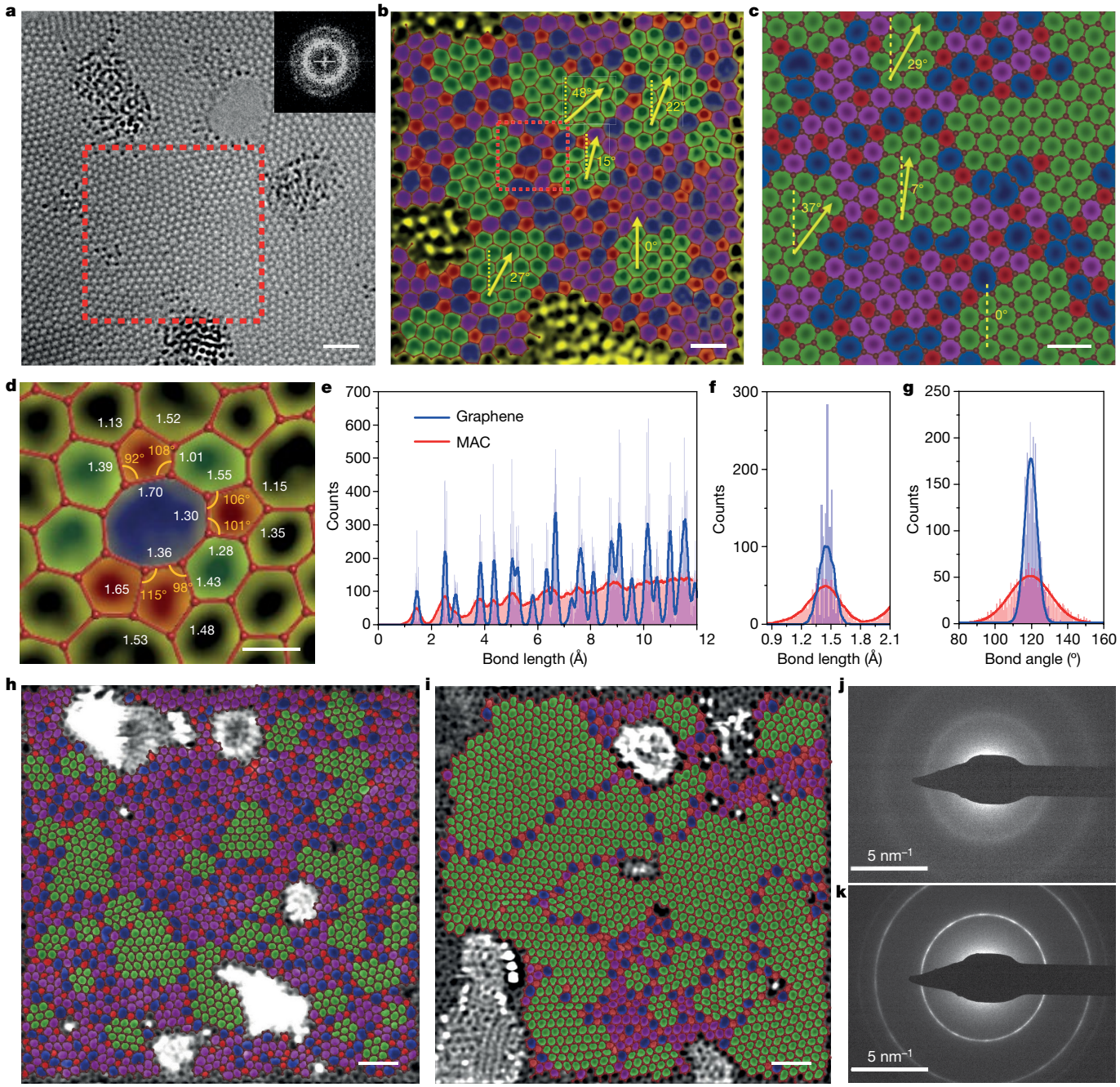


Fig. 2 | Atomic structure of MAC from TEM. a, Monochromated HRTEM image of MAC, 10 × 10 nm^2, and the corresponding Fourier transform of the selected region (inset). **b**, Large-scale atom-by-atom mapping of the selected region, 5 × 5 nm^2, in **a**. The image contrast is inverted and false-coloured for better visibility. Colour overlay is added for identification of pentagons (red), heptagons/octagons (blue) and strained hexagons (purple for individual hexagons; green for crystallite regions) that are omnipresent. Crystallites (regions of green hexagons) separate the regions with non-hexagons. Crystallites are defined to consist of at least a hexagonal ring surrounded by six hexagonal rings. Angular orientation of the hexagons changes across the image, as indicated by yellow arrows and the offset angle to the vertical. **c**, Theoretical model created as described in Extended Data Fig. 7, replicating the MAC features in **b** and displayed with same colour coding. **d**, Zoom-in region highlighted by the red square in **b**. The bond lengths (in ångström) and bond angles of each pentagon are precisely measured. **e**, Pair correlation function calculated by mapping the coordinates of each carbon atom in **b**. Graphene imaged under similar conditions with the same mapping algorithm is shown as a reference. **f**, Statistical histogram of the distribution of bond length to the first neighbouring atoms for MAC and graphene. **g**, Statistical histogram of the bond angle distribution for MAC and graphene. **h**, **i**, Comparison of scanning transmission electron microscopy (STEM) imaging for MAC (**h**) and nanocrystalline graphene (**i**), with false-colour overlay. Original images shown in Supplementary Note 3. **j**, **k**, SAED patterns over a 3-μm-diameter region: for the MAC sample shown in **h** (**j**) and for the nanocrystalline graphene sample shown in **i** (**k**). Scale bars, 1 nm (**a**), 0.5 nm (**b**, **c**), 0.2 nm (**d**), 1 nm (**h**, **i**) and 5 nm^{-1} (**j**, **k**).

(Fig. 2k). Dark-field TEM also reveals this, with only nanocrystalline graphene showing crystallinity (Extended Data Fig. 2 and Supplementary Note 3). To summarize, the direct imaging of the atomic structure clearly shows the amorphous nature of MAC to be (nano)crystallite, not Z-CRN.

The interlayer spacing of multilayer MAC was measured by atomic force microscopy (AFM) of overlapping MAC layers prepared by multiple wet transfers on a SiO_2/Si substrate. The spacing was found to be approximately 0.6 nm (Fig. 3b, Extended Data Fig. 3). This interlayer spacing is almost twice as large as that of graphene, similar to

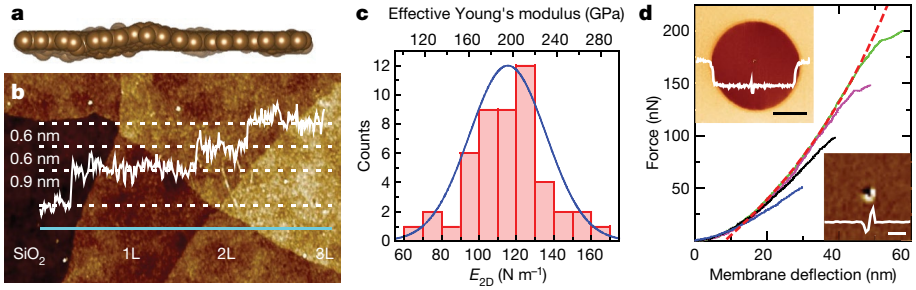


Fig. 3 | Mechanical properties of MAC. a, Side view of the model MAC shown from above in Fig. 2c, showing out-of-plane structural relaxation. **b**, AFM topography of one to three layers of MAC on SiO_2/Si. The cyan solid line indicates the line-scan position for the overlaid height profile. Image width, 3 µm. **c**, Histogram of the 2D elastic stiffness (E_{2D}) of 49 suspended MAC membranes with applied force of 100 nN, with an average of $E_{2D} = 115 \pm 21$ N m^{-1}. **d**, Selected force–displacement curves from multiple indentations with increasing force applied in 25-nN steps. Fitting to the linear elastic deformation expression is indicated by the dashed red line. Top inset, AFM scan of MAC suspended on a 2.5-µm-diameter well. The white line is the height profile along the centre with a 10.0-nm adhesion depth to the well wall. Bottom inset, close-up of collapsed MAC (the white line is the height profile along the centre) after repeated indentations at 200 nN. Scale bars, 1 µm (**d**, top inset) and 100 nm (**d**, bottom inset).

phosphorene's 0.6 nm, and consistent with the out-of-plane buckling caused by Stone–Wales defects[17]. AFM indentation experiments on suspended MAC membranes were used to determine the impact of structural buckling on its mechanical properties. The distribution of 2D elastic stiffness (E_{2D}) obtained by force–deflection measurements to 100 nN is centred at $E_{2D} = 115$ N m^{-1} (Fig. 3c). We observed an increasing slope (hence stiffness) of the force–displacement curves with consecutive indentations of increasing force (Fig. 3d). Hence, MAC undergoes plastic deformation with applied force. Furthermore, we obtain a breaking strength of 22 N m^{-1}, more than half the strength of single-crystal graphene[18]. We note that indentation rupture is restricted and does not propagate (Fig. 3c), in contrast to similar experiments with crystalline graphene that result in catastrophic failure by rapid crack propagation[18] (Supplementary Note 4).

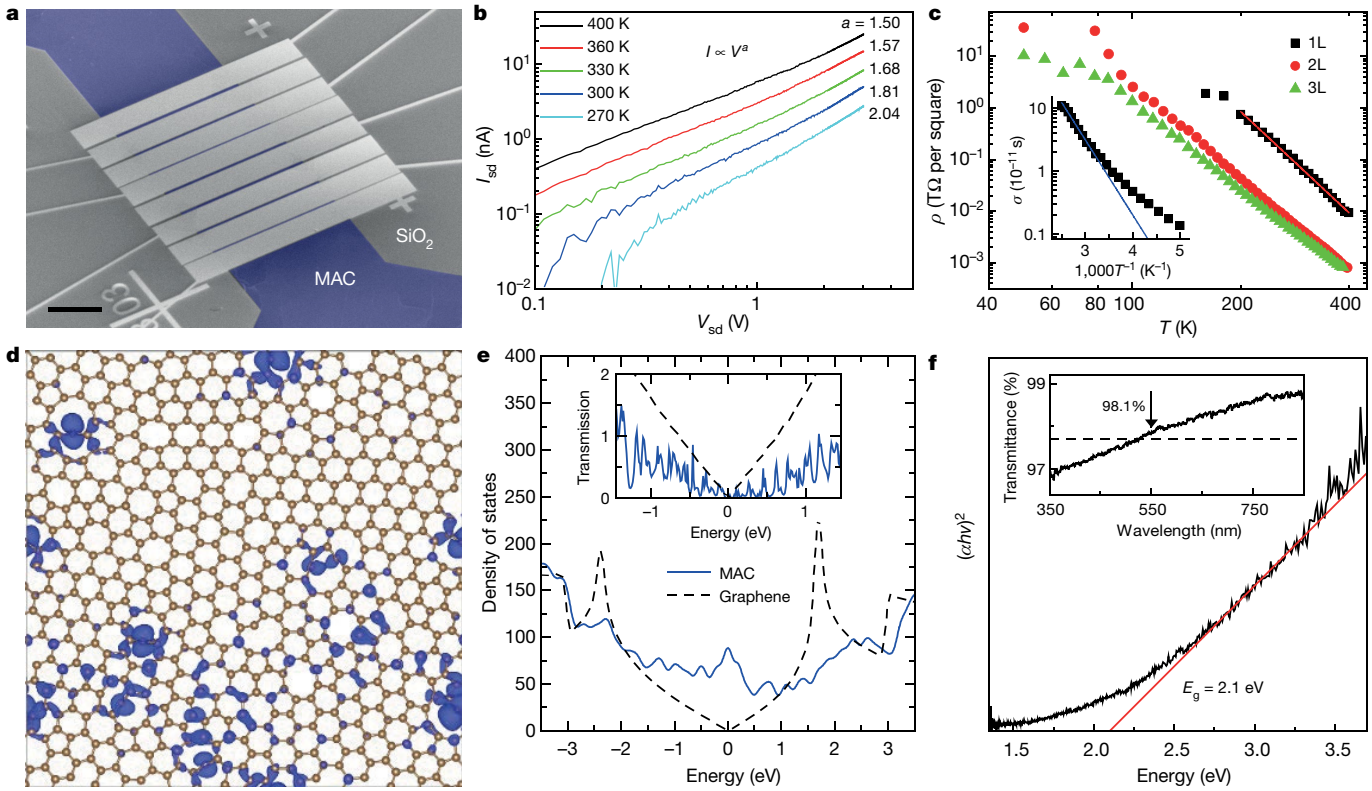


Fig. 4 | Insulating properties of MAC. a, False-colour SEM image of the two-terminal device with a channel width of 500 nm to 1 µm and electrodes 50 µm long. Scale bar, 20 µm. **b**, Nonlinear *I*–*V* curves measured at five different temperatures (Extended Data Fig. 4). **c**, Resistivity for samples with different layer (L) numbers as a function of temperature. The red line is the linear best fit with offset α and slope N equal to $10^{(26.9 \pm 0.05)}$ and -6.5 ± 0.05, respectively. Leakage current from SiO_2/Si prevents lower-temperature measurements. Inset, the Arrhenius plot gives an activation energy of 235 meV. **d**, Top view of the model MAC overlaid by the modulus square of the lowest unoccupied wavefunction. **e**, The density of states of the model MAC (solid blue line) shows a peak at the Fermi energy (zero energy on the plot) compared with that of graphene (dashed black line). Inset, phase-coherent electronic transmission through the model MAC (solid blue line) compared with that through graphene (dashed black line). **f**, Tauc plot calculated using an absorption coefficient α and photon energy $h\nu$ (h, Planck constant; ν, frequency) to determine the optical bandgap in amorphous materials, with extrapolation of the linear region (red line) estimating an optical bandgap of 2.1 eV. Inset, optical transmittance of MAC on CaF_2 substrate. The dashed line indicates pristine graphene absorption at 2.3%.

Next, we discuss source–drain bias-dependent electronic transport measurements (I–V) as a function of temperature and layer number (see also Extended Data Fig. 4). In all samples that we observed, $I = V^{\alpha}$, where $\alpha \approx 2$ (Fig. 4b) and the sheet resistance values are of order 100 GΩ at room temperature, similar to CVD boron nitride[19]. These sheet resistance values are high for a 2D material and are up to about 10 orders of magnitude larger than for graphene. Furthermore, the large activation energy of 235 meV (obtained from the Arrhenius plot in Fig. 4c) is similar to what has been observed in CVD boron nitride (193 meV)[20]. On the other hand, the temperature dependence of the resistivity ρ exhibits (Fig. 4c) a power-law dependence, $\rho = \alpha T^{N}$, instead of the expected Mott law ($\rho \propto e^{(T/T_0)^{-\gamma}}$, where α, N and γ are fitting factors) for variable-range hopping (Extended Data Fig. 5). In addition, we observe that the temperature dependence of all I–V curves collapses to a single curve, indicating an apparent power-law dependence (Extended Data Fig. 6). The two results together suggest that charge transport in MAC is unique in 2D and mimics that in disordered quasi-one-dimensional systems, known as rare-chain hopping[21,22].

A periodically repeatable structural model of MAC was constructed to resemble the observed structure in Fig. 2b, with similar distribution of ring size and similar randomly oriented crystallites embedded in a Z-CRN network (Fig. 2c; see Extended Data Fig. 7 for the construction process). Relaxation of the structure using density-functional-theory calculations produced buckling in the non-hexagon regions (Fig. 3a), as is known to occur at Stone–Wales defects in graphene[17]. We also constructed bilayer MAC by stacking the model structures at different relative displacements and minimized the total energy with respect to interlayer spacing. We obtained interlayer spacings ranging from 0.55 nm to 0.62 nm, in agreement with the measured interlayer spacings discussed above. Quantum molecular dynamics simulations of the structure at 300 K reveal that the MAC is stable and undergoes out-of-plane distortions in the entire structure in the form of long-wavelength flexural phonons, similar to those in graphene[23].

The experimental results described above are in good agreement with electronic-structure and quantum transport calculations performed with the MAC model described earlier (Fig. 2c). In Fig. 4e, we show the calculated MAC density of states (DOS) and compare it with that of crystalline graphene. It is clear that the V-shaped DOS of crystalline graphene around the Dirac point is no longer present. Instead, the DOS shows a relatively broad peak at the nominal Fermi energy, suggesting an increase of localized states[24]. In Fig. 4d, we plot the wavefunction (shown in blue) of a state within this peak, showing significant localization on non-hexagon rings that are separated by crystallites (Extended Data Fig. 8). The calculated transmission coefficient in the region around the Fermi energy, shown in the inset in Fig. 4e, approaches zero despite the high DOS, which explains the highly resistive charge transport.

To identify the effective bandgap of MAC, we measured its optical properties. The optical transmittance of MAC is 98.1% at 550 nm and increases towards 99% in the infrared (Fig. 4f). Clearly, graphene's 2.3% minimum absorption defined by the fine structure constant is no longer applicable. More importantly, a Tauc plot gives an optical bandgap of 2.1 eV with a long tail towards lower energies, which is reminiscent of 'tail states' in 3D amorphous silicon[25]. We also observed photoluminescence from MAC with a pronounced peak at 2.04 eV, which is not seen in graphene owing to the absence of a bandgap. Charge localization could be responsible for both the optical gap and the observation of photoluminescence, similar to insulating sp^3 amorphous carbon thin films with charge localization at isolated sp^2 clusters[26]. Finally, ellipsometry confirmed a bandgap at 2.1 eV and a corresponding relative permittivity of about 11 (Extended Data Fig. 9).

In summary, we have demonstrated a method for the growth of a stable amorphous material in the 2D limit, by producing large-area, freestanding monolayers of sp^2-bonded amorphous carbon films. We have furthermore demonstrated that the ring distribution is not a Z-CRN but resembles the competing crystallite model. Overall, so far, there exists no solid evidence for the formation of homogeneous, stable, freestanding Z-CRN monolayers, although their existence cannot be ruled out. Finally, stable MAC films directly grown on a wide range of surfaces at low temperatures are likely to be useful for a wide range of applications, including anti-corrosion barriers for heat-assisted magnetic recording on magnetic hard disks, and for current collectors and electrodes in batteries and supercapacitors. The possibility of engineering a distribution of carbon ring sizes also makes MAC attractive for proton transport membranes.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1871-2.

1. Zachariasen, W. H. The atomic arrangement in glass. *J. Am. Chem. Soc.* **54**, 3841–3851 (1932).
2. Voyles, P. M. & Abelson, J. R. Medium-range order in amorphous silicon measured by fluctuation electron microscopy. *Sol. Energy Mater. Sol. Cells* **78**, 85–113 (2003).
3. Gibson, J. M., Treacy, M. M. J., Sun, T. & Zaluzec, N. J. Substantial crystalline topology in amorphous silicon. *Phys. Rev. Lett.* **105**, 125504 (2010).
4. Treacy, M. M. J. & Borisenko, K. B. The local structure of amorphous silicon. *Science* **335**, 950–953 (2012).
5. Herman, I. P. Laser-assisted deposition of thin films from gas-phase and surface-adsorbed molecules. *Chem. Rev.* **89**, 1323–1357 (1989).
6. Wright, A. C. Neutron scattering from vitreous silica. V. The structure of vitreous silica: what have we learned from 60 years of diffraction studies? *J. Non-Cryst. Solids* **179**, 84–115 (1994).
7. Wright, A. C. The great crystallite versus random network controversy: a personal perspective. *Int. J. Appl. Glass Sci.* **5**, 31–56 (2014).
8. Roorda, S. & Lewis, L. J. Comment on "The Local Structure of Amorphous Silicon". *Science* **338**, 1539 (2012).
9. Eder, F. R., Kotakoski, J., Kaiser, U. & Meyer, J. C. A journey from order to disorder—atom by atom transformation from graphene to a 2D carbon glass. *Sci. Rep.* **4**, 4060 (2014).
10. Kotakoski, J., Krasheninnikov, A. V., Kaiser, U. & Meyer, J. C. From point defects in graphene to two-dimensional amorphous carbon. *Phys. Rev. Lett.* **106**, 105505 (2011).
11. Turchanin, A. et al. Conversion of self-assembled monolayers into nanocrystalline graphene: structure and electric transport. *ACS Nano* **5**, 3896–3904 (2011).
12. Joo, W.-J. et al. Realization of continuous Zachariasen carbon monolayer. *Sci. Adv.* **3**, e1601821 (2017).
13. Zandiatashbar, A. et al. Effect of defects on the intrinsic strength and stiffness of graphene. *Nat. Commun.* **5**, 3186 (2014).
14. Araujo, P. T., Terrones, M. & Dresselhaus, M. S. Defects and impurities in graphene-like materials. *Mater. Today* **15**, 98–109 (2012).
15. Zhang, C., Hao, X.-L., Wang, C.-X., Wei, N. & Rabczuk, T. Thermal conductivity of graphene nanoribbons under shear deformation: a molecular dynamics simulation. *Sci. Rep.* **7**, 41398 (2017).
16. Lichtenstein, L. et al. The atomic structure of a metal-supported vitreous thin silica film. *Angew. Chem. Int. Ed.* **51**, 404–407 (2012).
17. Lusk, M. T. & Carr, L. D. Nanoengineering defect structures on graphene. *Phys. Rev. Lett.* **100**, 175503 (2008).
18. Lee, C., Wei, X., Kysar, J. W. & Hone, J. Measurement of the elastic properties and intrinsic strength of monolayer graphene. *Science* **321**, 385–388 (2008).
19. Mahvash, F., Paradis, E., Drouin, D., Szkopek, T. & Siaj, M. Space-charge limited transport in large-area monolayer hexagonal boron nitride. *Nano Lett.* **15**, 2263–2268 (2015).
20. Kim, D. Y., Jeong, H., Kim, J., Han, N. & Kim, J. K. Defect-mediated in-plane electrical conduction in few-layer sp^2-hybridized boron nitrides. *ACS Appl. Mater. Interfaces* **10**, 17287–17294 (2018).
21. Rodin, A. S. & Fogler, M. M. Apparent power-law behavior of conductance in disordered quasi-one-dimensional systems. *Phys. Rev. Lett.* **105**, 106801 (2010).
22. Rodin, A. S. & Fogler, M. M. Hopping transport in systems of finite thickness or length. *Phys. Rev. B* **84**, 125447 (2011).
23. Morozov, S. V. et al. Giant intrinsic carrier mobilities in graphene and its bilayer. *Phys. Rev. Lett.* **100**, 016602 (2008).
24. Kapko, V., Drabold, D. A. & Thorpe, M. F. Electronic structure of a realistic model of amorphous graphene. *Phys. Status Solidi B* **247**, 1197–1200 (2010).
25. Tauc, J. Optical properties and electronic structure of amorphous Ge and Si. *Mater. Res. Bull.* **3**, 37–46 (1968).
26. Rusli Amaratunga, G. A. J. & Silva, S. R. P. Photoluminescence in amorphous carbon thin films and its relation to the microscopic properties. *Thin Solid Films* **270**, 160–164 (1995).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

MAC synthesis

MAC was synthesized by laser-assisted chemical vapour deposition (LCVD) process with krypton fluoride (KrF) pulsed laser (λ = 248 nm, Coherent Inc.). Copper foils (35 μm thick) were cleaned by sonication in acetone and isopropyl, followed by 30 min annealing in H_2 at 1,010 °C in a quartz tube furnace. Sputtered Au and Ru substrates were used without prior annealing. The growth substrates were transferred into the LCVD vacuum chamber and mounted on the titanium stage. The chamber was evacuated to base pressure of 10^{-6} mbar. CH_4 was introduced into the chamber to 10^{-2} mbar pressure. The substrate was exposed to a plasma (350 kHz pulsed d.c. generator at 20 W) and pulsed laser (40 mJ cm^{-2}, 50 Hz) over a 5-cm^2 area, for complete film formation under 1 min. To obtain larger area coverage, the stage was rastered with constant motion relative to the laser spot. Substrate temperatures were monitored by an infrared thermal sensor (Optis CTlaser 3MH_2-CF_4). Growth durations ranged from 1 min to 10 min, with temperatures due to laser heating between 200 °C and 500 °C; MAC growth was of similar quality. Substrate temperatures can be further controlled by a stage heater, for nanocrystalline graphene grown between 500 °C and 650 °C, and polycrystalline graphene grown at >650 °C. Although MAC synthesized on both sides of the copper foil has a similar amorphous quality, MAC from the underside (facing away from the laser) was used for detailed characterizations.

MAC sample preparation

MAC from the top side of the substrate was first removed by 5 min of 50-W Ar plasma. The Cu foil was wet etched in 0.7 wt% ammonium persulfate solution $(NH_4)_2S_2O_8$, by floating the growth substrate on the solution, with MAC facing up without a polymer support. It was transferred to deionized water for 2 h twice, using a rigid substrate, to rinse the interface (that is, it was twice transferred and rinsed) before scooping onto the desired substrate (SiO_2/Si wafers or PELCO holey SiN TEM grids with 2.5-μm pores). For transfer of MAC to SiO_2/Si substrate with holes for the indentation experiment, spin-coated PMMA (495 PMMA, A4, 4% in anisole, 4,000 rpm) was required as a polymer support during the transfer process. Spin-coated samples were baked on a hotplate at 180 °C for 2 min after spin-coating, and the same transfer steps were performed. Suspended MAC was dried supercritically from acetone in a critical point dryer system. The samples were annealed in 5% H_2 in Ar gas for 5 h to remove any organic residues.

Sample characterization

SEM images were acquired using a FEI Verios 460 field emission scanning microscope. Raman spectroscopy was measured by ALPHA 300 R from WITEC (532-nm excitation laser, 1-μm-diameter circular laser spot). Raman spectra for MAC grown on Cu and Au were measured after transfer to SiO_2 substrate, whereas the Raman spectrum was obtained directly on the Ru substrate. XPS spectra were measured using UHV Vacuum Generators ESCALAB Mk2 system with Omicron 7-Channeltron analyser, directly on the growth substrate for Cu, Au and Ru. The high-resolution angle-resolved XPS spectrum on Cu was measured using the Singapore Synchrotron Light Source at the National University of Singapore, with an Omicron EA 125 hemisphere energy analyser. XPS C 1*s* peaks were fitted after a Shirley background subtraction using XPSPEAK software with a 30% of Lorentzian–Gaussian ratio. The fitted data show a single sp^2-hybridized carbon peak with full-width half-maximum of 1.0. MAC was transferred to a calcium difluoride substrate for optical measurements. The ultraviolet–visible absorbance was recorded at room temperature on a SHIMADZU UV-3600 spectrophotometer. Photoluminescence measurements were performed with an NT-MDT NTEGRA SPECTRA confocal Raman microscope in backscattering geometry. We used the excitation laser at 473 nm and a 100× objective lens with numerical aperture of 0.6 to focus to a spot size of approximately 1.5 μm. The laser power density was kept below 1 kW cm^{-2} to avoid heating effects. The photoluminescence signal was detected with a charge-coupled device array operating at −80 °C. Optical measurements were performed in high vacuum (10^{-5} mbar).

Transmission electron microscopy

The dark-field TEM (DF-TEM) imaging was conducted in an FEI Tecnai F30 microscope operating at 80 kV. DF-TEM images were recorded using the diffraction rings selected by the objective aperture, with a recording time of 120 s for each image. MAC HRTEM imaging was performed on a non-commercialized JEOL ARM60 microscope equipped with a Schottky field emission gun, a JEOL double Wien filter monochromator, double delta correctors and a Gatan OneView camera with high stability. A slit of 2.8 μm was used for energy filtering, enabling a spatial resolution of 1.1 Å with a beam current of about 20 pA. The microscope was operated at 60 kV. The STEM imaging was done in a JEOL 2100F with delta probe corrector, which corrects the aberration up to fifth order, resulting in a probe size of 1.0 Å. The imaging was conducted at an acceleration voltage of 60 kV. The convergent angle for illumination was about 35 mrad, with a collection detector angle ranging from 45 mrad to 200 mrad. A JEOL heating-holder was used in all experiments to heat the sample up to 700 °C during imaging. The pair correlation function and bond angle distribution were calculated from the coordinates of the carbon atoms determined in the HRTEM and STEM images by a home-built atom-finding algorithm.

Computational method

The theoretical model for the crystallite MAC was constructed using the kinetic Monte Carlo method[27]. The simulation starts with a supercell, 40 × 40 Å^2, containing 610 carbon atoms randomly placed in the x–y plane. The density of carbon atoms is the same as in crystalline graphene. The supercell is then relaxed using a conjugated gradient algorithm. The interatom interactions are described using the adaptive intermolecular reactive empirical bond-order (AIREBO) potential by Stuart[28] as implemented in the LAMMPS molecular dynamics program[29]. Starting from the initial configuration, kinetic Monte Carlo is used for annealing. In each Monte Carlo iteration, one Stone–Wales transformation is performed to the current configuration: that is, one randomly chosen bonded carbon–carbon pair is rotated by 90° in the x–y plane. The probability to accept the new relaxed structure is min $\{1, \exp((E_{old} - E_{new})/k_B T)\}$, where E_{old} is the energy of the current configuration, E_{new} is the energy of the new configuration after relaxation, and $k_B T$ is selected to be 0.5 eV (k_B, Boltzmann constant). We perform more than 60,000 iterations and generate 1,560 distinct configurations. We select the configuration at the 20,000th iteration as the theoretical model for its similarity with the experimental HRTEM image (Extended Data Fig. 7). The electronic structure is calculated using density functional theory in the generalized gradient approximation[30] as implemented in the VASP software[31]. The coherent electronic transmission probability is calculated using the Transiesta programme.

Mechanical characterization

AFM topography was measured by tapping mode using a Bruker AFM. This was also used for indentation experiments. MAC was suspended on wells with diameter of 1–2.5 μm. Force–displacement curves were obtained from indentation experiments performed at the centre of the suspended membranes using single-crystal diamond probes with 8-nm diamond tips (spring constant of 1–5 N m^{-1}) attached to silicon cantilevers (K-TEK nanotechnology). The applied force was then increased until fracture to obtain the breaking strength of MAC. The 2D elastic constant of MAC was calculated by fitting the force–displacement results using the equation for force[32]

$$F = \frac{2\pi\sigma\delta}{\ln(a/r)} + \frac{Eq^3\delta^3}{a^2} \quad (1)$$

where a and r are the radius of the suspended membrane and the AFM tip, respectively. σ, δ and E are the 2D pre-tension, deflection and 2D effective Young's modulus of the membrane, respectively, and q (~1.02) is a function of Poisson's ratio (taken as 0.165).

The breaking strength corresponding to a breaking force F_b was then calculated as

$$\sigma^{2D} = \left(\frac{F_b E}{4\pi r}\right)^{1/2} \quad (2)$$

Electronic transport characterization

Multi-terminal devices of one-, two- and three-layer-thick MAC were fabricated on a Si/SiO_2 wafer using the standard microfabrication procedures of electron beam lithography. Oxygen plasma was first used to pattern MAC to fixed width of 50 µm. Devices were created with channel length between 200 nm and 1,000 nm. The width-to-length ratio ranged from 50 to 250. The geometry of each sample is shown in Supplementary Table. After the second step of lithography, a set of Co electrodes (35 nm thick) were deposited in an ultrahigh vacuum system by an electron-beam evaporator at the deposition pressure of 1×10^{-8} torr. Contacts 10 µm wide were made to reduce the contact resistance (Fig. 4a). Electrical measurements were carried out in the d.c. regime using two Keithley 2400 Sourcemeters for source–drain and gate voltages, which allows applying a voltage and measure the current simultaneously. The gate leakage was lower than 1 nA at room temperature and below 25 pA at 150 K. All measurements, for 2D amorphous carbon films, were performed in helium-free cryostats and in vacuum better than 10^{-5} mbar to ensure reproducible measurement conditions. Resistivity measurements are performed at source-drain voltage of 1 V. Resistance was extracted from the I–V characteristic using a polynomial fit and taken at the first derivative at zero voltage.

Data availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Code availability

The code used in this study is available upon request from the corresponding author.

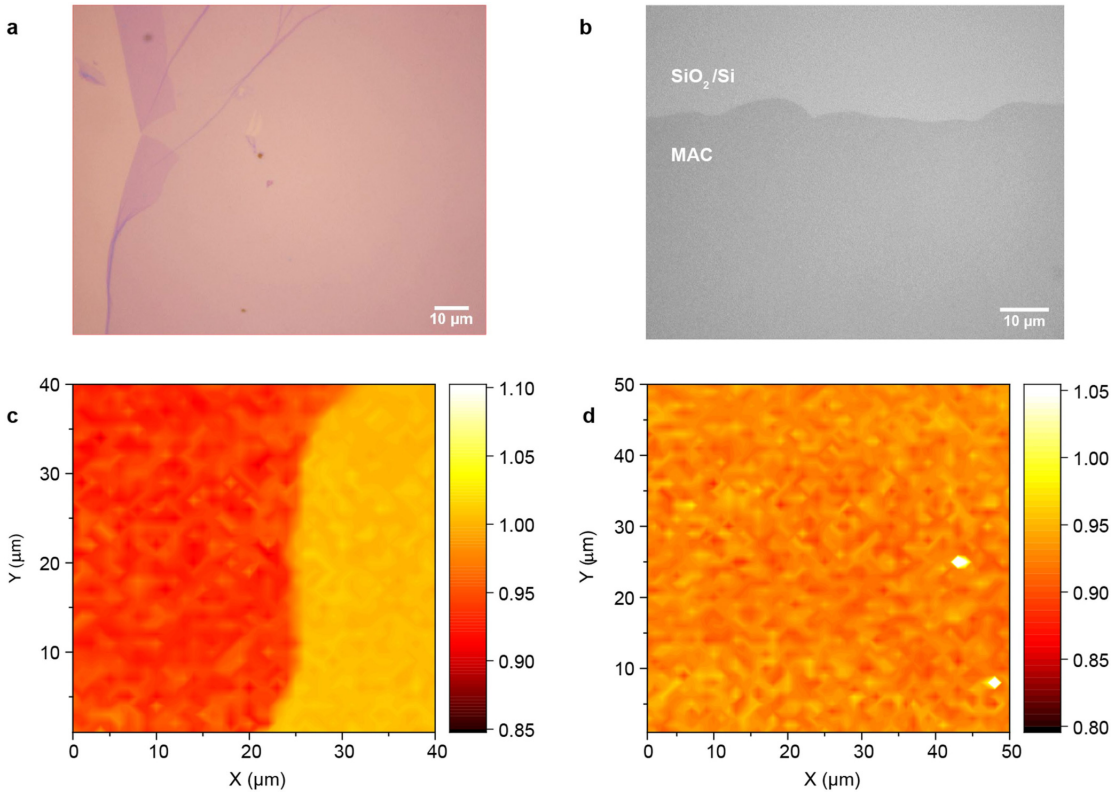
27. Zhuang, J., Zhao, R., Dong, J., Yan, T. & Ding, F. Evolution of domains and grain boundaries in graphene: a kinetic Monte Carlo simulation. *Phys. Chem. Chem. Phys*. **18**, 2932–2939 (2016).
28. Stuart, S. J., Tutein, A. B. & Harrison, J. A. A reactive potential for hydrocarbons with intermolecular interactions. *J. Chem. Phys*. **112**, 6472–6486 (2000).
29. Plimpton, S. Fast parallel algorithms for short-range molecular dynamics. *J. Comput. Phys*. **117**, 1–19 (1995).
30. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett*. **77**, 3865–3868 (1996).
31. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
32. Ruiz-Vargas, C. S. et al. Softened elastic response and unzipping in chemical vapor deposition graphene membranes. *Nano Lett*. **11**, 2259–2263 (2011).

Acknowledgements B.Ö. acknowledges support by the National Research Foundation, Prime Minister's Office, Singapore, under its Competitive Research Programme (CRP award number NRF-CRP9-2011-3), NRF Investigatorship (NRFI award number NRF-NRFI2018-08) and Medium-Sized Centre Programme. Work at Vanderbilt University (H.S., Y.W. and S.T.P.) was supported by US Department of Energy grant DE-PG02-09ER46554 and by the McMinn Endowment. Computations were performed at the National Energy Research Scientific Computing Center, a DOE Office of Science User Facility supported by the Office of Science of the US Department of Energy under contract number DE-AC02-05CH11231, and at the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by US National Science Foundation grant number ACI-1053575. J.L. and K.S. acknowledge JSPS KAKENHI (JP16H06333 and P16823) for financial support. This work was also supported by the Pico Center at MCPC of SUSTech that receives support from Presidential fund and Development and Reform Commission of Shenzhen Municipality. We thank X. Yu and Singapore Synchrotron Light Source (SSLS, SINS beamline) for assistance with angle-resolved XPS measurements, D. Vella and G. Eda for assistance with photoluminescence measurements and G. K. W. Koon for assistance with ellipsometry measurements. B.Ö. thanks J. Fabian for discussions.

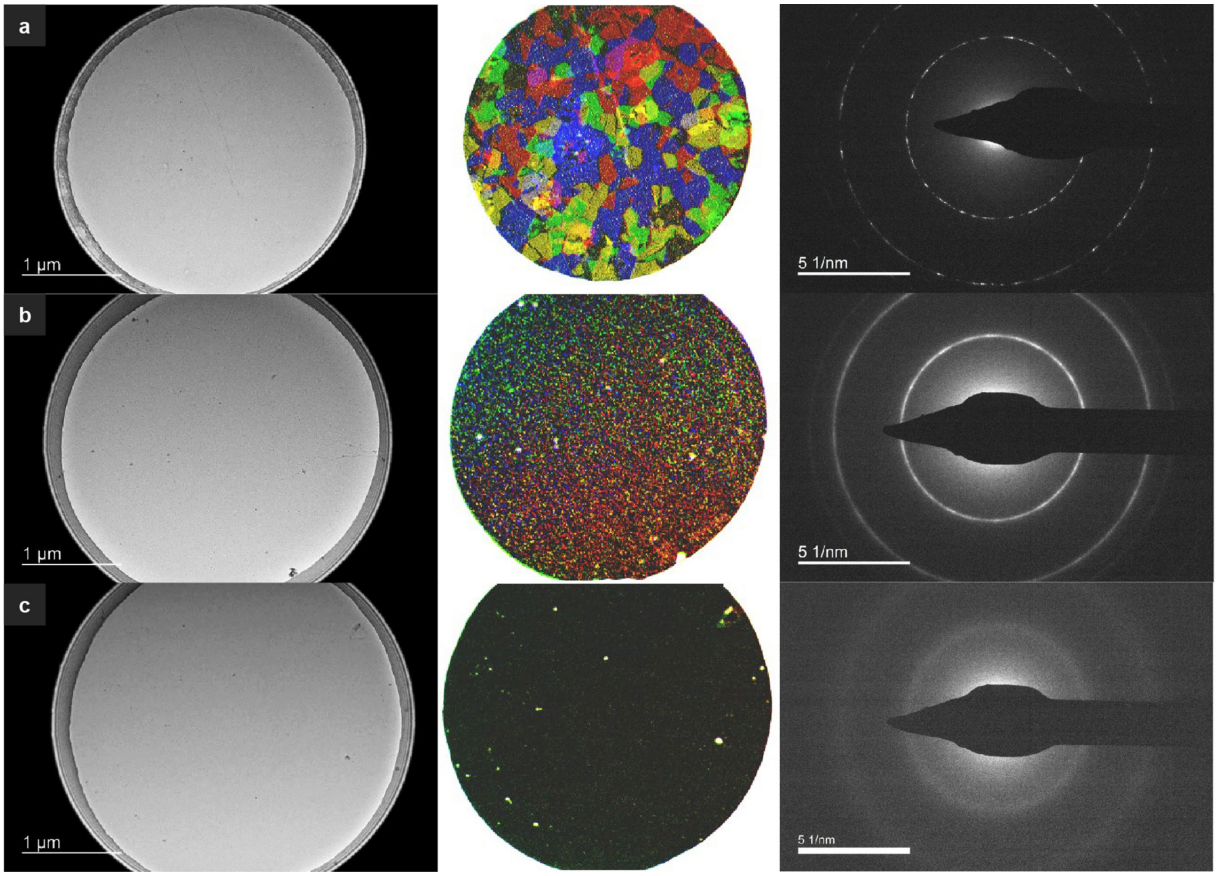
Author contributions B.Ö. initiated and coordinated the work. C.-T.T. and H.Z. contributed equally to the work. C.-T.T., C.O. and B.Ö. designed the experiments. H.Z., D.B.F., C.O., C.-T.T., N.K. and I.H.A. synthesized the material. H.Z., D.B.F. and C.O. performed the Raman, XPS and AFM characterizations. J.L., Z.G. and K.S. performed the DF-TEM, HR-TEM and STEM characterizations. H.S., Y.P.W. and S.T.P. performed the density functional theory calculations and Monte Carlo simulations. D.B.F. and H.A. fabricated the device and performed the indentation experiment. A.S.M. fabricated the device and performed the electrical measurements. C.-T.T. performed the optical measurements. C.-T.T., S.T.P. and B.Ö. analysed all data and wrote the paper. All authors discussed and commented on the manuscript.

Competing interests The authors declare no competing interests.

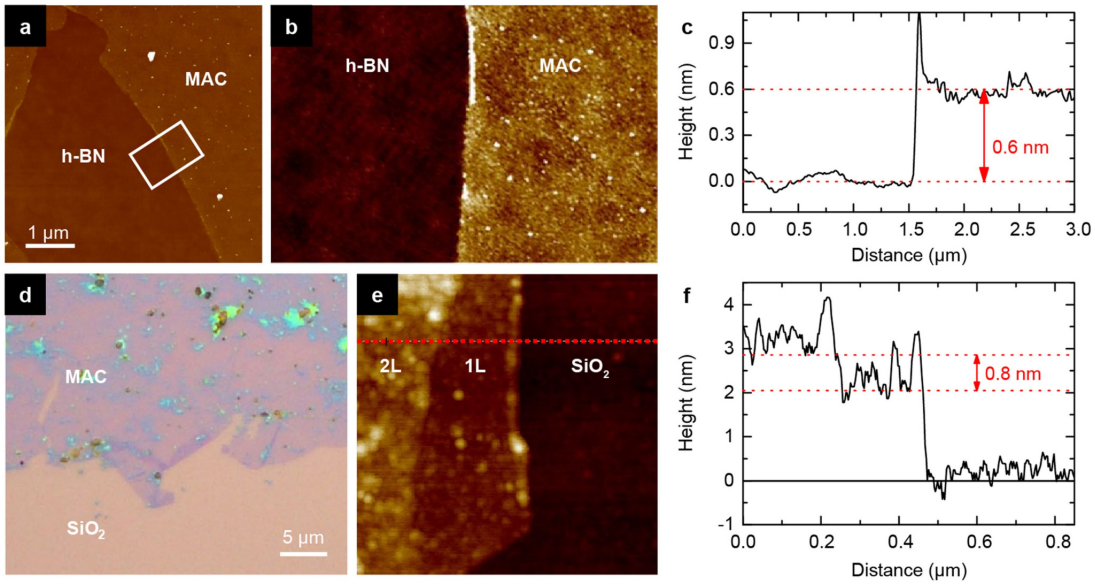
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1871-2.
Correspondence and requests for materials should be addressed to B.Ö.
Peer review information *Nature* thanks Volker Deringer, Jannik Meyer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



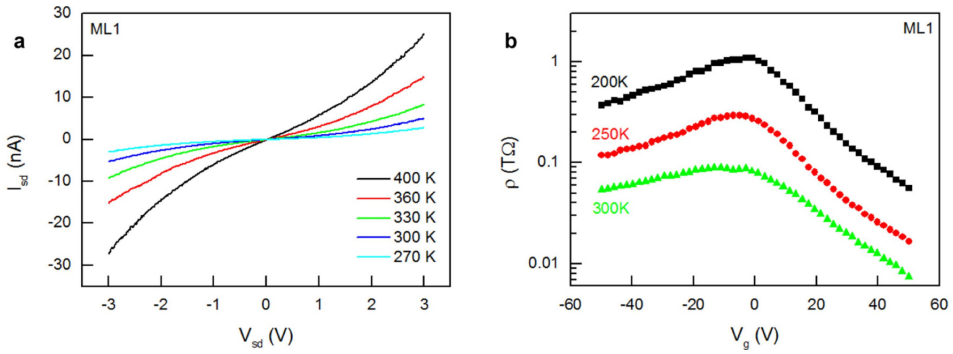
Extended Data Fig. 1 | Optical image and Raman mapping of MAC film transferred on Si/SiO2 substrate. a, MAC transferred to SiO_2/Si wafer. Crease and fold at MAC edge from transfer process. **b**, Contrast-enhanced image in greyscale shows no topological features visible on MAC. **c**, Raman mapping shows that the intensity ratio of the D to G band, I_D/I_G, is uniform at the edge of transferred MAC on a Si/SiO_2 substrate. This ratio is slightly higher than the expected ratio of 0.82, as the Raman background signal is not subtracted during Raman mapping. **d**, Raman map (50 × 50 μm^2, with 2,500 data points) of I_D/I_G shows uniform quality over a large area (standard deviation is ±1.8%).



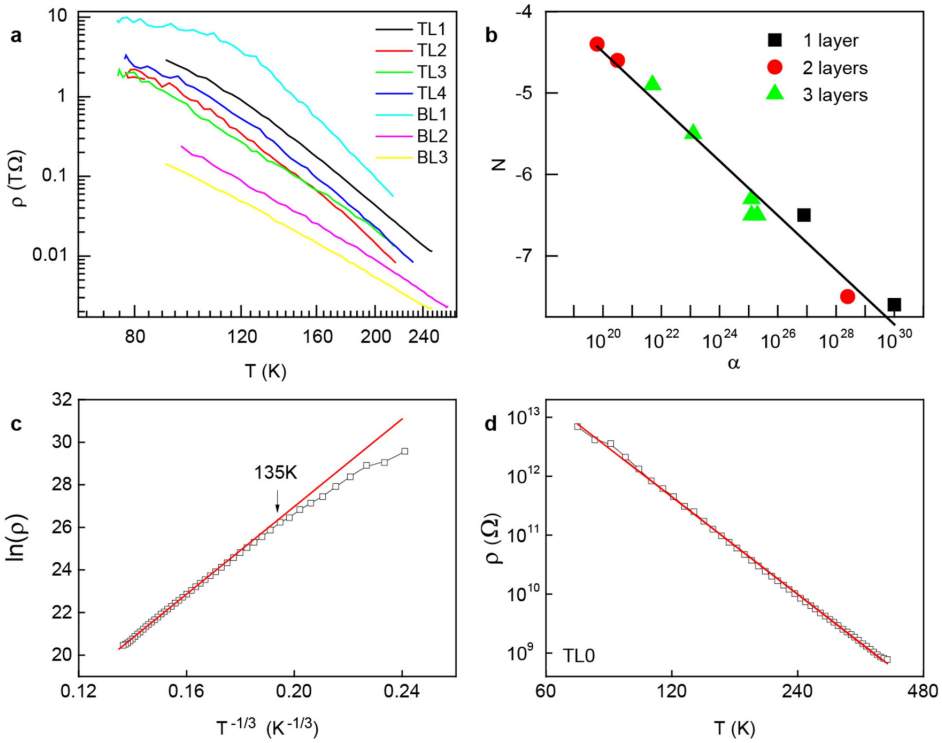
Extended Data Fig. 2 | Comparison of large-area TEM data. a, Polycrystalline graphene with grain size 200–500 nm. **b**, Nanocrystalline graphene (sample in Fig. 2i) with grain size 1–3 nm. **c**, MAC (sample in Fig. 2h). Left, bright-field TEM; centre, DF-TEM with false-colour image overlay showing crystal domains; right, SAED patterns. Both polycrystalline and nanocrystalline graphene show crystal domains under DF-TEM and have well defined SAED diffraction patterns. MAC has no visible domains in DF-TEM. Only MAC has the characteristic amorphous halo from SAED diffraction.



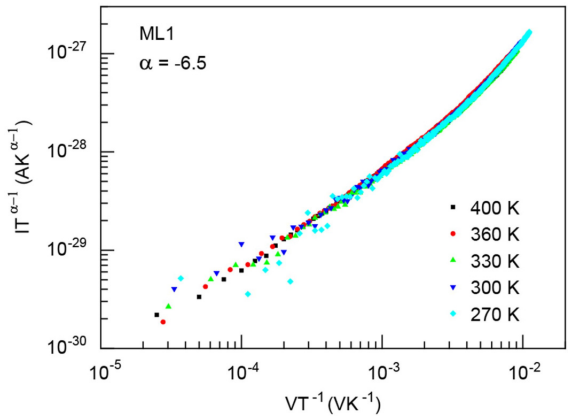
Extended Data Fig. 3 | Thickness of MAC. a, AFM topography of MAC transferred onto an atomically flat, exfoliated BN crystal. **b**, Close-up of the region indicated by the white box in **a**, with a scan length of 3 μm. **c**, Height profile of MAC, obtained by averaging vertically over the entire image scan in **b**, shows a thickness of about 0.6 nm. **d**, Optical image of MAC grown on gold substrates and transferred onto SiO_2 by wet transfer with KI/I_2 solution as gold etchant and followed by 7 h annealing in Ar/H_2 at 300 °C. For MAC transferred from gold, residues from the gold etching cannot be fully removed owing to the non-optimized transfer process. **e**, AFM topography of a folded MAC edge in **d** to obtain one and two layers. **f**, The height profile along the red dotted line in **e** gives the thickness of MAC grown on gold as about 0.8 nm. Residual contamination from the transfer step and difference in substrate interactions may account for the higher thickness of MAC when transferred from gold.



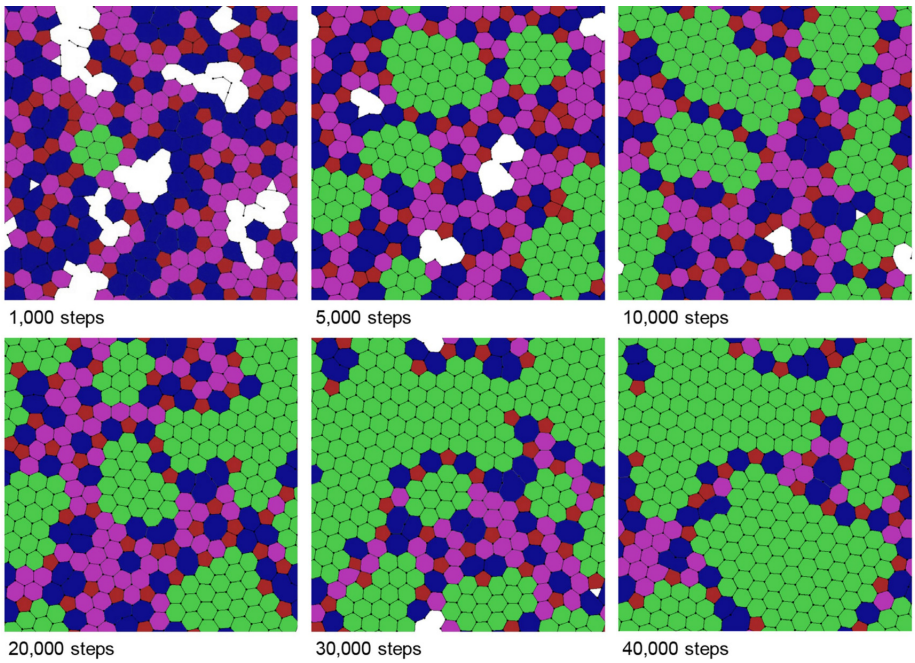
Extended Data Fig. 4 | Current–voltage measurement and gate dependence. **a**, Nonlinear *I*–*V* curves in Fig. 4b shown in linear scale. **b**, Gate dependence of MAC with electrostatic gating by SiO_2/Si back gate shows ambipolar behaviour and strong temperature dependence, increasing from 100 GΩ per square at room temperature to 1 TΩ per square at 200 K.



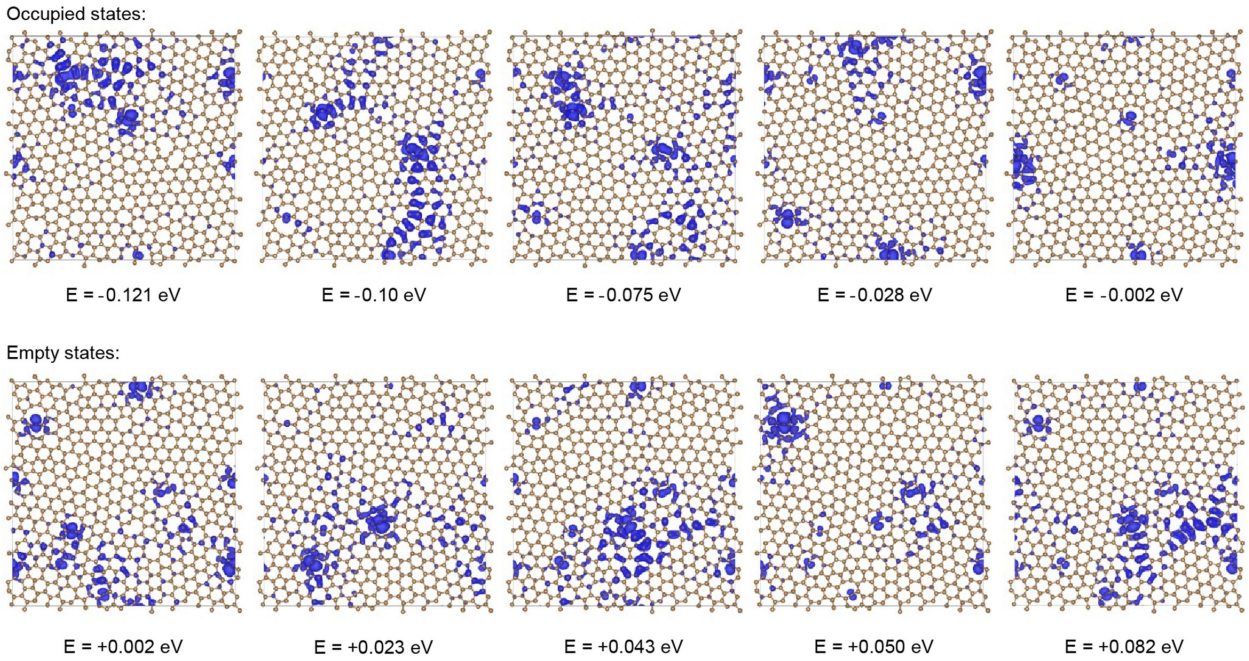
Extended Data Fig. 5 | Resistivity of samples as a function of the temperature. a, Temperature dependence of the resistivity for a set of devices with two- and three-layer MAC (denoted BL and TL, respectively). **b**, The linear fit of the resistivity of offset α and power N, as described by the formula $\rho = \alpha T^N$, shows a linear correlation, $N = 2.2(\pm 0.6) - 0.33(\pm 0.02)\log\alpha$. The data in **a**, **b** are provided in Supplementary Table. **c**, **d**, Resistivity of tri-layer MAC sample (sample ID TL0) as a function of temperature plotted in a linear scale. The Mott plot (**c**) shows a temperature limit of 135 K (indicated by the black arrow) up to which it can be applied, as opposed to a power-law dependence (**d**). The red lines are the best linear fits.



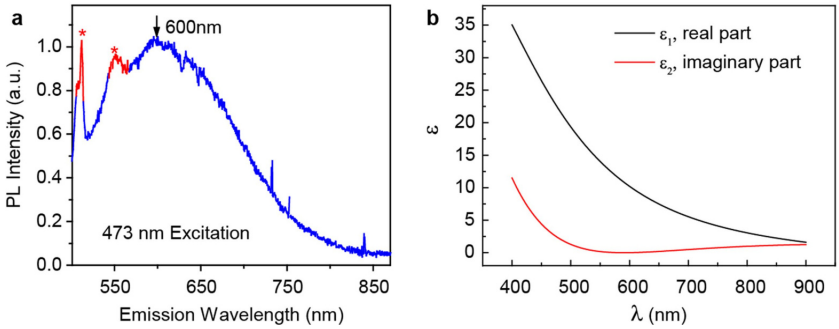
Extended Data Fig. 6 | Universal scaling of nonlinear *I*–*V* characteristics for ML1 sample. All *I*–*V* curves are collapsed to a single curve indicating apparent power-law dependence in a disordered quasi-one-dimensional system. The equation for the universal scaling curve is from ref. [21].



Extended Data Fig. 7 | Selection of MAC theoretical model. The configuration for the MAC structure at the 20,000th iteration is chosen as the theoretical model, shown in Fig. 2c, for its similarity to experimental HRTEM images of MAC. Previous steps also show an unstable ring structure (for example, rings with fewer than four carbon atoms, or ill-defined rings). Furthermore, the ring size distribution from 10,000 steps is similar . Colour overlay is added for identification of pentagons (red), heptagons/octagons (blue) and strained hexagons (purple or green) that are omnipresent. Crystallites (green) separate regions with non-hexagons. Crystallites are defined to consist of at least one hexagonal ring surrounded by six hexagonal rings.



Extended Data Fig. 8 | Modulus square of electronic wavefunctions. The lowest unoccupied wavefunctions (blue) are shown for ten states with energies closest to the Fermi energy. This shows localization on non-hexagon rings that are separated by crystallites.



Extended Data Fig. 9 | Optical measurements of MAC on SiO_2/Si. a, Photoluminescence spectra were measured with an excitation wavelength of 473 nm. The peaks marked by asterisks correspond to the Raman signal from MAC. The emission spectrum of the film has a broad peak centred at 600 nm. **b**, Relative permittivity of MAC on SiO_2/Si substrate by ellipsometry measurements. E_g is taken where the imaginary part equals zero, at about 590 nm (2.1 eV), at which the real part of the relative permittivity is approximately 11.

Article

Proton-assisted growth of ultra-flat graphene films

https://doi.org/10.1038/s41586-019-1870-3

Received: 4 December 2018

Accepted: 30 September 2019

Published online: 8 January 2020

Guowen Yuan[1,2], Dongjing Lin[1,2], Yong Wang[1,2], Xianlei Huang[1,2], Wang Chen[1,2], Xuedong Xie[1,2], Junyu Zong[1,2], Qian-Qian Yuan[1,2], Hang Zheng[1,2], Di Wang[1,2], Jie Xu[1,2], Shao-Chun Li[1,2], Yi Zhang[1,2], Jian Sun[1,2], Xiaoxiang Xi[1,2] & Libo Gao[1,2]*

Graphene films grown by chemical vapour deposition have unusual physical and chemical properties that offer promise for applications such as flexible electronics and high-frequency transistors[1–10]. However, wrinkles invariably form during growth because of the strong coupling to the substrate, and these limit the large-scale homogeneity of the film[1–4,11,12]. Here we develop a proton-assisted method of chemical vapour deposition to grow ultra-flat graphene films that are wrinkle-free. Our method of proton penetration[13–17] and recombination to form hydrogen can also reduce the wrinkles formed during traditional chemical vapour deposition of graphene. Some of the wrinkles disappear entirely, owing to the decoupling of van der Waals interactions and possibly an increase in distance from the growth surface. The electronic band structure of the as-grown graphene films shows a V-shaped Dirac cone and a linear dispersion relation within the atomic plane or across an atomic step, confirming the decoupling from the substrate. The ultra-flat nature of the graphene films ensures that their surfaces are easy to clean after a wet transfer process. A robust quantum Hall effect appears even at room temperature in a device with a linewidth of 100 micrometres. Graphene films grown by proton-assisted chemical vapour deposition should largely retain their intrinsic performance, and our method should be easily generalizable to other nanomaterials for strain and doping engineering.

Wafer-sized graphene films can be easily produced by chemical vapour deposition (CVD)[1,7]. However, the performance of CVD graphene is still inferior to that of small flakes exfoliated from bulk graphite[18]. The grain boundaries[5], atomic vacancies and impurities[19] inside CVD-grown graphene are considered to be bottlenecks to their homogeneous quality, although some methods have been reported to overcome these challenges[3,8]. Besides this, electrons in graphene are weakly localized in two dimensions with a ballistic transport range over micrometres[20], so not only grain boundaries, vacancies and impurities but also wrinkles can act as scattering centres. Wrinkles in graphene are considered a kind of linear defect[11] that will degrade the large-scale physical and chemical properties. Therefore, both perfect atomic lattices and flatness across a large scale will be equally important for super-high performance[21].

During the CVD process, monolayer graphene should adhere to the surface of a growth substrate, such as copper[1,7,8] or platinum[3]. The mismatch in thermal expansion coefficients between graphene and the substrate causes wrinkles (illustrated in Fig. 1a)[4,22] as an effective way to release compressive strain. These are easily formed in rough areas such as atomic steps or surface vacancies[12]. Some approaches have been reported to reduce wrinkles, such as using low thermal expansion coefficients[3], single-crystalline substrates[9] or growth at low temperature[7], but the wrinkles have seemed impossible to eliminate entirely. Trying to reduce the interaction between graphene and the growth substrate may be the key to this bottleneck[2], as illustrated in Fig. 1b.

CVD processes usually need high temperature (>800 °C) to decompose a gaseous carbon source[3,8], and there are plenty of atomic hydrogens, protons and electrons in the reaction atmosphere. The graphene is impermeable to atomic hydrogen, but the hydrogen can transform it into graphane (C–H bonds) at room temperature (RT)[23,24]. The protons are able to penetrate graphene and hexagonal boron nitride (h-BN) by an electrochemical process[14–16] or as energetic particles[17,25]. Once they have penetrated into the material, the protons may recombine with electrons to form atomic hydrogen again[25], bonded to the substrates or existing in other forms. The intercalation of hydrogen is assumed to be capable of decoupling graphene from the substrates after annealing in an H_2 atmosphere[2,26,27], but the thermal process will not produce enough atomic hydrogen and is not effective for delamination[28]. Therefore, for complete decoupling of graphene films, more protons are needed. In this study of graphene films on Cu substrates, we controllably generate protons by using an inductively coupled plasma (ICP) of hydrogen[13], instead of the sole thermal process (Fig. 2a).

We have simulated theoretically the bonding status of recombined hydrogen atoms that lie between the graphene and Cu, for different temperatures and different densities of atomic hydrogen. First-principles molecular dynamics evolution shows that the hydrogen atoms will move towards the Cu surface. Some of them bond to Cu atoms, some diffuse to interstitial positions between the topmost Cu layers, and some re-bond into hydrogen molecules (H_2), which can move freely and

[1]National Laboratory of Solid State Microstructures and School of Physics, Nanjing University, Nanjing, China. [2]Collaborative Innovation Center of Advanced Microstructures, Nanjing University, Nanjing, China. *e-mail: lbgao@nju.edu.cn

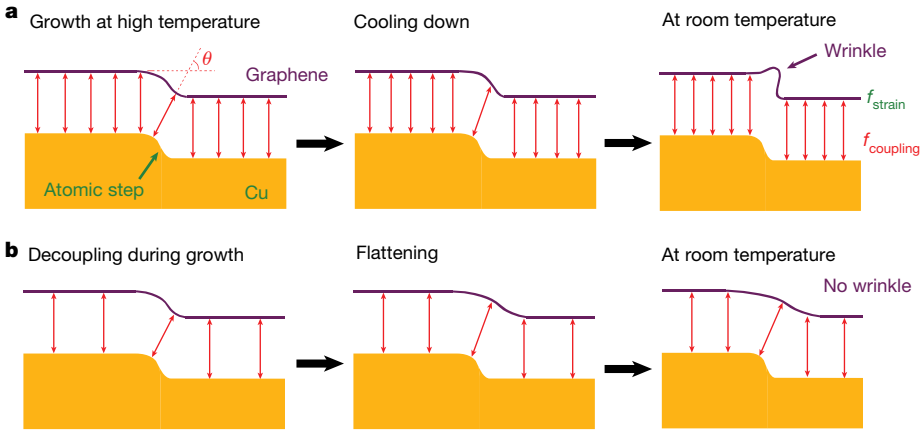


Fig. 1 | Illustration of graphene wrinkle formation and avoidance. a, Schematic illustration of wrinkle formation at an atomic step (or rough area) of a shrinking Cu surface when cooling down. **b**, Schematic illustration of the disappearance of wrinkle when decoupling graphene from its substrate during CVD growth.

influence the van der Waals interaction between graphene and the Cu substrate. At equilibrium, the simulations at RT result in a distance of ~3.9 Å between the Cu surface and the graphene for H:C = 1:2 and ~3.6 Å for H:C = 1:4. This surface distance changes to ~6.4 Å for H:C = 1:2 and ~3.6 Å for H:C = 1:4 at 400 °C (Fig. 2b). Therefore, the higher density of atomic hydrogen is useful to decouple the interlayer interactions by increasing the distance between the substrate surface and the graphene layer, and the higher reaction temperature also appears to be helpful to the formation of H_2 (more details are given in Extended Data Fig. 1a, b and Supplementary Video 1).

Based on the theoretical results, we monitor the evolution of wrinkles under different ICP treatments. First, graphene films are grown on single-crystalline Cu(111) films by a traditional CVD process. Wrinkles invariably exist in these films (Extended Data Fig. 1c–e), despite growth at lower temperature (Extended Data Fig. 2a) or with a higher H:C ratio (Extended Data Fig. 2b, c). Atomic force microscope (AFM) imaging of as-grown graphene films on Cu(111) shows sharp atomic steps on the surface (Fig. 2c). The wrinkles (bright lines) occur along or perpendicular to the Cu(111) atomic steps, indicating the existence of the graphene films. There are also numerous nano-sized bubbles (bright dots). The AFM phase image appears homogeneous except for the regions of wrinkles, bubbles and atomic steps (inset of Fig. 2c).

This wrinkled graphene film is then treated by ICP in H_2 atmosphere at 400 °C and 650 °C in turn (Fig. 2d, e). The ex situ AFM images show obvious changes in wrinkles and bubbles. After ICP at 650 °C, most wrinkles are reduced, and some have been eliminated. Meanwhile, more bubbles arise, and they show major differences in the AFM phase images. After ultrahigh vacuum (UHV, $<10^{-6}$ Pa) annealing at 400 °C, the bubbles are enlarged, but more wrinkles have disappeared completely (Fig. 2f). The enlarging and increasing bubbles indicate the existence of H_2 molecules[17], formed by rebonding of atomic hydrogen. Height profiles across the same wrinkle after different stages of treatment are shown in Fig. 2g, marked by pink lines in Fig. 2c–f. This wrinkle in the as-grown film is about 4 nm in height, but it reduces to ~3.2 nm after ICP at 400 °C, disappears after ICP at 650 °C and never returns even after UHV annealing.

Typical Raman spectra of graphene films before and after ICP treatment and UHV annealing are plotted in Fig. 2h. There are no obvious D bands after the ICP treatments, demonstrating that no graphane or atomic defects were formed by this controllable plasma at a temperature of more than 300 °C (more details are given in Extended Data Fig. 3a)[23]. After ICP treatment at 400 °C, the Raman frequencies of the G (ω_G) and 2D (ω_{2D}) bands show no obvious shift, but their relative intensity of I_{2D}/I_G becomes a little smaller. Both ω_G and ω_{2D} show significant blueshift after ICP treatment at 650 °C and UHV annealing, which results from the increased compressive strain in graphene[12,22,29,30]. These changes can be reproduced on weakly defective graphene films on Cu(111) and micromechanical exfoliated monolayer graphene on SiO_2/Si, but cannot be reproduced on heavily defective graphene because of large atomic vacancies, and cannot be reproduced by helium ICP, for which there are no protons in the reaction gas (Extended Data Figs. 3b–d and 4). All of this indicates that the van der Waals interactions between graphene and the underlying surface have been changed by the penetrating protons and recombined hydrogens, but not by the direct diffusion of molecular hydrogen.

Although post-treatment by ICP can effectively reduce the density and height of wrinkles once they have formed, there are still some wrinkles, along with newly generated bubbles. The schematic illustration in Extended Data Fig. 5 is shown to make clear that these wrinkles cannot be removed. Large-area Raman mapping also shows that the distribution of ω_{2D} is not homogeneous (Extended Data Fig. 6a, b). We therefore developed the CVD growth process further by supplying enough protons by creating a plasma in an H_2 atmosphere, and hence avoiding the formation of wrinkles in the first place. The numerous protons that have penetrated into the material lead to re-bonded H_2 lying between the as-grown graphene and Cu(111), and this is intended to decouple the van der Waals interactions during the CVD growth so that the grown graphene can slide freely without being dragged by the atomic steps or atomic vacancies. A photo of 4-inch graphene films grown on Cu(111) by proton assistance is shown in Fig. 3a: it looks homogeneous on the large scale, and a typical AFM image of this surface is shown in the inset. The roughness of the ICP–CVD graphene films is compared with that of CVD graphene on Cu foil and on a Cu(111) substrate in Fig. 3b (details in Extended Data Fig. 6c). This developed growth process retains the roughness of the Cu(111) surface, and no sharp atomic steps appear after growth, which is different from the traditional CVD process. The ω_G and ω_{2D} of multi-point Raman spectra collected from 30 μm × 30 μm regions are plotted in Fig. 3c (details in Extended Data Fig. 6d). ICP–CVD graphene films appear more compressed than the others, and the Raman frequencies appear similar to those of wrinkled graphene films after ICP treatment at 650 °C. The higher compressive strains of these ICP–CVD graphene films are consistent with the absence of wrinkles. Similar to protons, deuterons can also permeate freely into graphene[15], and this could also be used to flatten wrinkled graphene by ICP post-treatment to grow ultra-flat graphene films (Extended Data Fig. 6e, f).

Moiré patterns, formed by the strong van der Waals interactions, can easily be observed in the graphene/Cu[9] and graphene/SiC[26] systems. However, there are no obvious moiré patterns in the ICP–CVD graphene films (inset of Fig. 3d and Extended Data Fig. 7a). Scanning tunnelling spectroscopy (STS) spectra were collected at different points marked in the scanning tunnelling microscope (STM) images in Fig. 3d. They

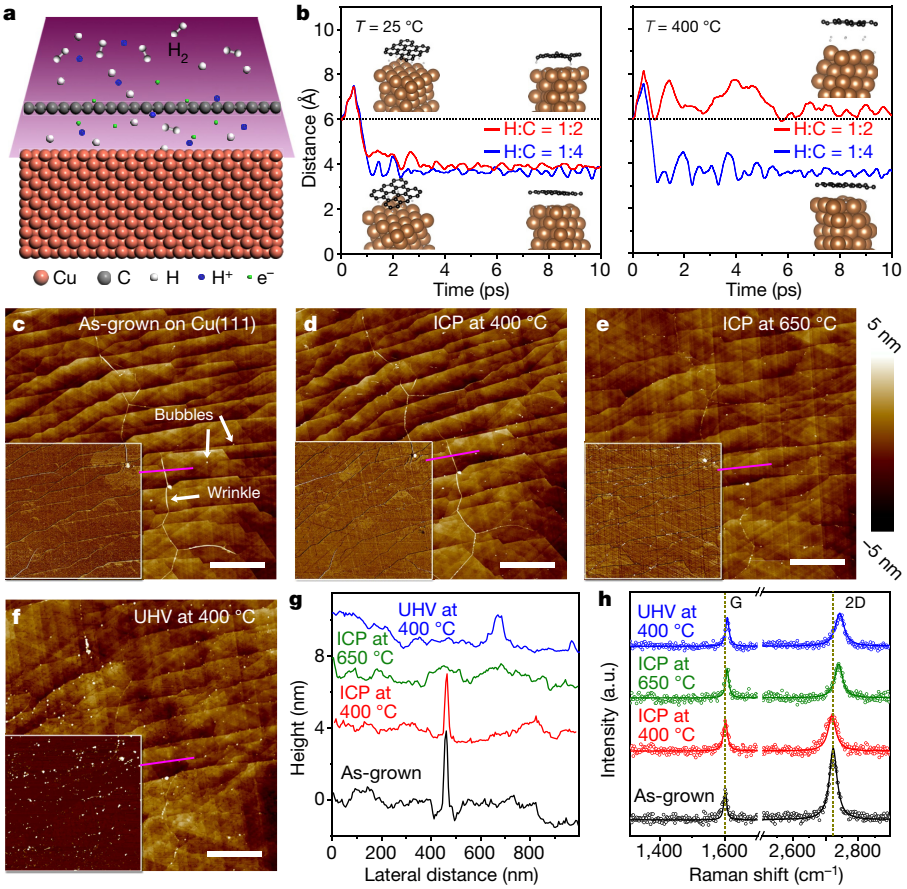


Fig. 2 | Proton-assisted relaxation of graphene wrinkles on copper. a, Illustration of protons penetrating through the as-grown graphene film during ICP treatment. The van der Waals interaction between the graphene and the substrate is altered by the recombined hydrogen (H atoms and H_2). **b**, Theoretical simulation of bonding status for recombined hydrogen at different density. A high density will lead to the production of re-bonded H_2, the presence of which can decouple the interaction between the graphene and substrate, at 400 °C. **c**–**f**, Typical AFM images of wrinkled graphene film on Cu(111) after different treatments: the surface as grown by CVD (**c**); after ICP at 400 °C (**d**); after ICP at 650 °C (**e**); and after UHV annealing at 400 °C (**f**). Insets are the AFM phase images. The wrinkles and bubbles in the graphene films are changed during the ICP treatments. **g**, Height profiles for the same wrinkle, taken along the line marked in pink in **c**–**f**. The wrinkle disappears completely after ICP at 650 °C. **h**, Typical Raman spectra of graphene films as-grown and after ICP treatment, showing that no graphane or atomic defects arise. All the AFM images use the same height bar of −5 nm to 5 nm, and the length scale bars in **c**–**f** are all 1 μm.

all show no energy gaps and obvious V-shape Dirac cones around the Fermi levels. Linear dispersion features are homogeneously present at the atomic planes and steps on the Cu(111) surface. Moreover, the moiré patterns and energy gaps in the wrinkled graphene/Cu(111) system could be almost eliminated by the ICP post-treatment (Extended Data Fig. 7b, c). Angle-resolved photoemission spectroscopy (ARPES) is also performed to characterize these samples. At the K point (Fig. 3e), linear conical Dirac dispersion appears in the ICP–CVD films. Their π-bands intersect at a Dirac point of approximately 40 meV below the Fermi level, indicating slight n-type doping. Graphene suspended across a micro-hole shows a Dirac point at 0 V without any doping. The feature of the slight shift of the Dirac point in the sample in Fig. 3e is similar to those in suspended samples. However, the quasi-suspended feature is in a metastable state, and it can easily be destroyed by annealing at higher temperature (Extended Data Fig. 7a, d). The Dirac point of this wrinkled film on Cu(111) before ICP treatment declines at about 310 meV, and a mini bandgap of about 130 meV emerges at the Dirac point, formed by the strong coupling interactions between graphene and Cu[31]. After ICP treatment at 650 °C, the Dirac point can rise to about 170 meV and the bandgap reduce to about 90 meV, implying that the coupling between graphene and the substrates becomes weaker. It is worth noting that the doping levels of the ICP post-treated graphene films are also inhomogeneous, and they are easily changed during the UHV annealing procedure (Extended Data Fig. 7e, f).

For better monitoring of the van der Waals interactions between graphene and substrates, we performed in situ Raman measurements with variable temperature (T) for different graphene films. In principle, if the graphene films have the same interlayer coupling to the substrates, their Raman frequencies (ω) will show the same $\Delta\omega/\Delta T$ for different graphene films because of the same compressive strain change ($\Delta\varepsilon/\Delta T$). After temperature cycling, the Raman spectrum returns to its previous form at the same temperature, demonstrating that the temperature cycling process is non-destructive (Extended Data Fig. 8a). When the temperature decreases from 300 K to 10 K, ω_G and ω_{2D} both show a blueshift (Fig. 4a). Their shifts are approximately proportional to the temperature variation, and the relative intensity ratio and full-width at half-maximum (FWHM) barely change during this temperature cycling (Extended Data Fig. 8b). The $\Delta\omega_G/\Delta T$ and $\Delta\omega_{2D}/\Delta T$ are estimated to be 2.65×10^{-2} cm^{-1} K^{-1} and 7.31×10^{-2} cm^{-1} K^{-1} for the as-grown wrinkled films, and they decrease to approximately 1.82×10^{-2} cm^{-1} K^{-1} and 4.05×10^{-2} cm^{-1} K^{-1} after ICP treatment at 650 °C. The reduced $\Delta\omega/\Delta T$ indicates an insensitive $\Delta\varepsilon/\Delta T$ and a gradual decoupling during these ICP treatments. Moreover, the $\Delta\omega_G/\Delta T$ and $\Delta\omega_{2D}/\Delta T$ for ICP–CVD grown graphene are approximately 1.70×10^{-2} cm^{-1} K^{-1} and 4.20×10^{-2} cm^{-1} K^{-1}, indicating that ICP treatment during or after CVD growth has a similar effect on decoupling graphene.

Pure graphene/SiO_2/Si and graphene/Cu(111) without hydrogen treatment (hydrogen ICP treatment or annealing in hydrogen atmosphere)

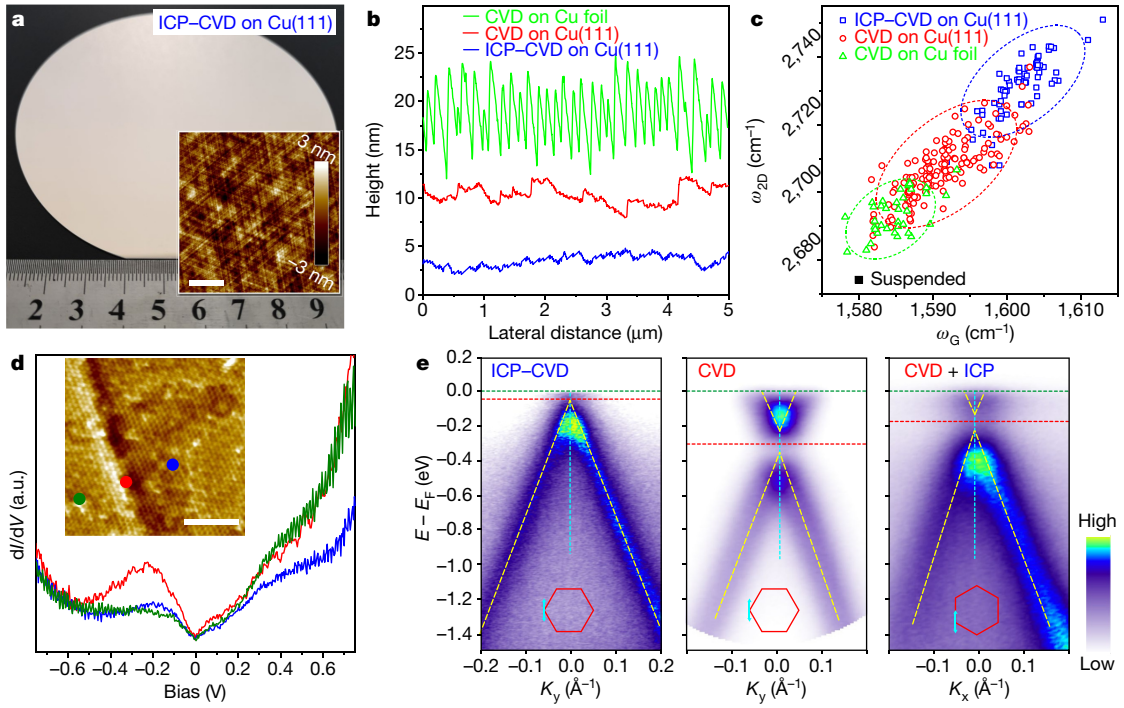


Fig. 3 | Proton-assisted growth of wrinkle-free and quasi-suspending graphene films. a, Photo of 4-inch graphene film grown on Cu(111). Inset, typical AFM image. **b**, Typical smoothness of substrates after graphene growth: Cu foil and Cu(111) covered by wrinkled graphene films, and Cu(111) covered by ICP–CVD grown graphene films. **c**, Distribution of ω_G and ω_{2D} obtained from multiple points in the ICP–CVD grown films and other wrinkled samples, along with a suspending film (black solid) as reference. **d**, STS spectra on ICP–CVD grown graphene film across a Cu(111) atomic step. Their dispersions appear linear (that is, V-shaped rather than U-shaped) and there is no obvious doping. Inset, the related STM image. **e**, ARPES of as-grown graphene film on Cu(111) with proton assistance, and wrinkled samples before and after ICP treatment, showing that their doping level can be changed by proton penetration. The green lines on red hexagons show the ARPES cutting orientations and locations in the Brillouin zone of graphene. Scale bars: inset to **a**, 2 μm; inset to **d**, 2 nm, respectively.

before and after UHV annealing are also investigated for comparison. In the transferred graphene/SiO_2/Si, the surface distance is reduced by about 5 Å after UHV annealing at 430 °C, and this will cause a stronger van der Waals interaction ($\Delta\varepsilon/\Delta T$) and larger $\Delta\omega/\Delta T$ (Extended Data Fig. 8c)[21,29]. In the pure graphene/Cu(111) system, UHV annealing should also be an efficient way to change their coupling through adjusting their surface distance. The ω_G and ω_{2D} are changed after UHV annealing at higher than 200 °C, and all of these changes are irreversible. The $\Delta\omega_G$ and $\Delta\omega_{2D}$ extracted for graphene transferred to Cu(111), obtained from variable-temperature Raman measurements, are plotted in Fig. 4b (details in Extended Data Fig. 8d). The $\Delta\omega_G/\Delta T$ and $\Delta\omega_{2D}/\Delta T$ before UHV annealing are approximately 1.67×10^{-2} cm^{-1} K^{-1} and 4.11×10^{-2} cm^{-1}K^{-1}, and these change to approximately 2.22×10^{-2} cm^{-1}K^{-1} and 5.97×10^{-2} cm^{-1} K^{-1} after UHV at 430 °C. It is notable that these $\Delta\omega_G/\Delta T$ and $\Delta\omega_{2D}/\Delta T$ for as-transferred graphene have the same values as those of ICP–CVD graphene and wrinkled graphene after ICP post-treatment. It is well known that the as-transferred graphene films before annealing have weak van der Waals coupling to the substrates. Hence, the penetrating protons and recombined hydrogens may reduce their coupling level by increasing their surface distance, which is consistent with the theoretical simulations above.

The ultra-flat nature makes it easier to remove the surface residues that are inevitably introduced by the transfer procedures[10]. Figure 4c counts the number and size of residual nanoparticles (NPs) on different graphene films. The average numbers are about 40, 100 and 400 for ultra-flat films, wrinkled films on Cu(111) and Cu foils, respectively, in regions 30 μm × 30 μm. Typical AFM images of these transferred graphene films are shown in the inset of Fig. 4c and Extended Data Fig. 9. There are almost no NPs larger than 10 nm in the ultra-flat graphene films, whereas other films usually have larger and denser residual NPs. The inherent wrinkles will adsorb large-sized NPs, which are more difficult to remove completely.

The ultra-flat and easy-clean surface of the graphene strongly indicates homogeneous performance. Here, the robustness of the quantum Hall effects (QHE) is measured. We fabricated standard Hall bar devices with 100 μm linewidth, for which the carrier mobility detected from the field effect extracted from longitudinal resistance (R_{xx}) is approximately 4,900 cm^2 V^{-1} s^{-1} at RT (Extended Data Fig. 10a)[6]. Figure 4d shows R_{xx} and Hall conductivity (σ_{xy}) as a function of gate voltage (V_{bg}) under a fixed magnetic field of 7.5 T at different temperatures. The first σ_{xy} plateaus, which are a hallmark of the Hall effect, appear at $2e^2/h$ for both electrons and holes even at RT[32]. At 1.5 K, the fourth integer plateau is already visible, indicating the extraordinary transport properties of this sample. The robust QHE in the ultra-flat films indicates ballistic transport properties, and the performances of all of these graphene films should be much closer to their extraordinary intrinsic nature than seen in wrinkled films (more in Extended Data Fig. 10b–d).

In summary, a proton-assisted CVD method has been developed to grow ultra-flat graphene films by decoupling their van der Waals interaction with the Cu substrates. Their ultra-flat nature has been demonstrated by the strong compressive strain, quasi-suspending status, decoupled interaction and wrinkle-free morphology. Furthermore, the decoupling process originates from the protons permeating into the graphene and the recombined hydrogens that then result, rather than the direct intercalation of molecular hydrogen. The ultra-flatness could allow graphene films to retain more homogeneous properties on a large scale, such as their easy-clean surface and robust QHE, and will keep their properties closer to the intrinsic values than is the case for wrinkled films. This proton penetration process should be easily generalized to other materials to tune their doping or strains.

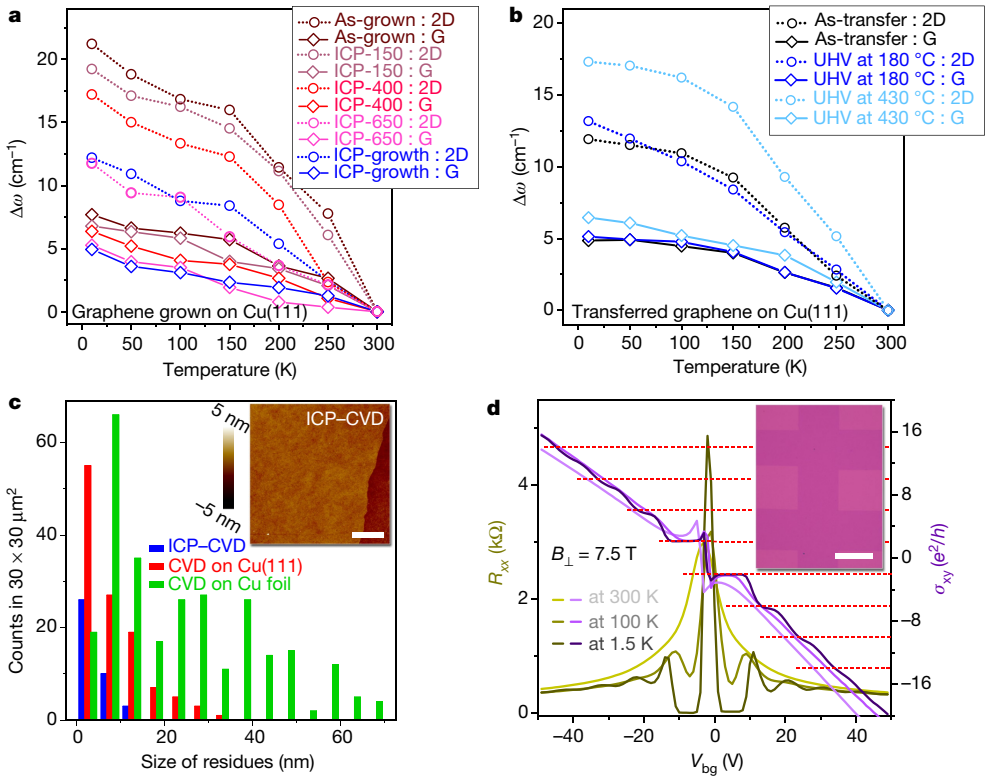


Fig. 4 | Easy-clean nature and robust QHE in ultra-flat graphene films. a, $\Delta\omega_G$ and $\Delta\omega_{2D}$ extracted from in situ Raman measurements of wrinkled graphene before and after ICP treatments, and the ICP–CVD grown graphene as reference. The temperature is varied from 10 K to 300 K. The results indicate that the $\Delta\varepsilon/\Delta T$ values become less sensitive to temperature change after higher-temperature ICP treatment (150, 400, 650 indicate temperatures in K for the ICP process). **b**, Extracted $\Delta\omega_G$ and $\Delta\omega_{2D}$ for graphene transferred onto Cu(111), before and after UHV annealing; $\Delta\varepsilon/\Delta T$ values become sensitive to change of temperature after UHV annealing at higher temperature. **c**, Residual NPs on transferred graphene films after standard wet transfer procedures and UHV annealing. Insets, typical AFM images for these ultra-flat films. **d**, Robust QHE and Hall plateaus can be easily observed for the large-scale, ultra-flat graphene films in a magnetic field $B = 7.5$ T at RT. Inset, optical image of the Hall bar device. Scale bars: inset to **c**, 1 μm; inset to **d**, 100 μm.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1870-3.

1. Bae, S. K. et al. Roll-to-roll production of 30-inch graphene films for transparent electrodes. *Nat. Nanotechnol.* **5**, 574–578 (2010).
2. Lee, J. H. et al. Wafer-scale growth of single-crystal monolayer graphene on reusable hydrogen-terminated germanium. *Science* **344**, 286–289 (2014).
3. Gao, L. B. et al. Repeated growth and bubbling transfer of graphene with millimetre-size single-crystal grains using platinum. *Nat. Commun.* **3**, 699 (2012).
4. Chae, S. J. et al. Synthesis of large-area graphene layers on poly-nickel substrate by chemical vapor deposition: wrinkle formation. *Adv. Mater.* **21**, 2328–2333 (2009).
5. Huang, P. Y. et al. Grains and grain boundaries in single-layer graphene atomic patchwork quilts. *Nature* **469**, 389–392 (2011).
6. Petrone, N. et al. Chemical vapor deposition-derived graphene with electrical performance of exfoliated graphene. *Nano Lett.* **12**, 2751–2756 (2012).
7. Gao, L. B. et al. Face-to-face transfer of wafer-scale graphene films. *Nature* **505**, 190–194 (2014).
8. Wu, T. R. et al. Fast growth of inch-sized single-crystalline graphene from a controlled single nucleus on Cu–Ni alloys. *Nat. Mater.* **15**, 43–47 (2016).
9. Deng, B. et al. Wrinkle-free single-crystal graphene wafer grown on strain-engineered substrates. *ACS Nano* **11**, 12337–12345 (2017).
10. Zhang, Z. K. et al. Rosin-enabled ultraclean and damage-free transfer of graphene for large-area flexible organic light-emitting diodes. *Nat. Commun.* **8**, 14560 (2017).
11. Zhu, W. et al. Structure and electronic transport in graphene wrinkles. *Nano Lett.* **12**, 3431–3436 (2012).
12. Bronsgeest, M. S. et al. Strain relaxation in CVD graphene: wrinkling with shear lag. *Nano Lett.* **15**, 5098–5104 (2015).
13. Mao, M. & Bogaerts, A. Investigating the plasma chemistry for the synthesis of carbon nanotubes/nanofibres in an inductively coupled plasma enhanced CVD system: the effect of different gas mixtures. *J. Phys. D* **43**, 205201 (2010).
14. Hu, S. et al. Proton transport through one-atom-thick crystals. *Nature* **516**, 227–230 (2014).
15. Lozada-Hidalgo, M. et al. Sieving hydrogen isotopes through two-dimensional crystals. *Science* **351**, 68–70 (2016).
16. Mertens, S. F. L. et al. Switching stiction and adhesion of a liquid on a solid. *Nature* **534**, 676–679 (2016).
17. He, L. et al. Isolating hydrogen in hexagonal boron nitride bubbles by a plasma treatment. *Nat. Commun.* **10**, 2815 (2019).
18. Novoselov, K. S. et al. A roadmap for graphene. *Nature* **490**, 192–200 (2012).
19. Ishikawa, R. et al. Direct electric field imaging of graphene defects. *Nat. Commun.* **9**, 3878 (2018).
20. Morozov, S. V. et al. Strong suppression of weak localization in graphene. *Phys. Rev. Lett.* **97**, 016801 (2006).
21. Dean, C. R. et al. Boron nitride substrates for high-quality graphene electronics. *Nat. Nanotechnol.* **5**, 722–726 (2010).
22. Cheong, H., Yoon, D. & Son, Y. W. Negative thermal expansion coefficient of graphene measured by Raman spectroscopy. *Nano Lett.* **11**, 3227–3231 (2011).
23. Elias, D. C. et al. Control of graphene's properties by reversible hydrogenation: evidence for graphane. *Science* **323**, 610–613 (2009).
24. Guisinger, N. P., Rutter, G. M., Crain, J. N., First, P. N. & Stroscio, J. A. Exposure of epitaxial graphene on SiC(0001) to atomic hydrogen. *Nano Lett.* **9**, 1462–1466 (2009).
25. Stolyarova, E. et al. Observation of graphene bubbles and effective mass transport under graphene films. *Nano Lett.* **9**, 332–337 (2009).
26. Riedl, C., Coletti, C., Iwasaki, T., Zakharov, A. A. & Starke, U. Quasi-free-standing epitaxial graphene on SiC obtained by hydrogen intercalation. *Phys. Rev. Lett.* **103**, 246804 (2009).
27. Ahn, S. J. et al. Electronic structure of graphene grown on a hydrogen-terminated Ge (110) wafer. *J. Korean Phys. Soc.* **73**, 656–660 (2018).
28. Kim, J. et al. Layer-resolved graphene transfer via engineered strain layers. *Science* **342**, 833–836 (2013).
29. Lee, J. E., Ahn, G., Shim, J., Lee, Y. S. & Ryu, S. Optical separation of mechanical strain from charge doping in graphene. *Nat. Commun.* **3**, 1024 (2012).
30. Kiraly, B. et al. Electronic and mechanical properties of graphene-germanium interfaces grown by chemical vapor deposition. *Nano Lett.* **15**, 7414–7420 (2015).
31. Avila, J. et al. Exploring electronic structure of one-atom thick polycrystalline graphene films: a nano angle resolved photoemission study. *Sci. Rep.* **3**, 2439 (2013).
32. Novoselov, K. S. et al. Room-temperature quantum Hall effect in graphene. *Science* **315**, 1379 (2007).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

CVD growth of graphene

Proton-assisted growth of graphene films is performed in a customized ICP–CVD system. The typical growth parameters are as follows: sputtered 800-nm Cu–Ni(111) films on *c*-plane sapphire, growth temperature of 650 °C, pressure of 6 Pa, CH_4/H_2 ratio of 1:20, plasma power of 15 W and growth time of 5 min. Traditional CVD-grown graphene films are produced in a tube furnace, with typical growth parameters as follows: sputtered 800-nm Cu–Ni(111) films on *c*-plane sapphire or commercial Cu foils, growth temperature of 1,050 °C, atmospheric pressure, CH_4/H_2/Ar ratio of 0.02:10:500 and growth time of 30 min. We use Cu–Ni alloy (90% Cu + 10% Ni) instead of Cu in this study to avoid sublimation at 1,050 °C or UHV annealing at 650 °C, and maintain the smooth surface after growth. To simplify writing, we use Cu(111) instead of Cu–Ni(111) alloy in the main text.

Transfer of CVD graphene

After growth, graphene/Cu(111)/sapphire or graphene/Cu foil is spin-coated with polymethyl methacrylate (PMMA) (molecular weight 996 kD, 4 wt% in ethyl lactate, 2,000 r.p.m. for 1 min). For the samples used for variable-temperature Raman measurement, a thermal release tape is used to cover the baked PMMA/graphene/Cu(111)/sapphire. An aqueous solution of 1 M $(NH_4)_2S_2O_8$ is used to etch Cu, and the tape/PMMA/graphene is pulled out and cleaned with DI water three times. After drying the graphene at 80 °C for 12 h, tape/PMMA/graphene is pasted on SiO_2/Si or Cu(111) substrate, then baked at 100 °C to bond graphene/Cu(111) tightly. All the pasting processes are operated in a glovebox to avoid Cu oxidation. Finally, PMMA is completely removed by acetone. For the samples used for transport measurements: first, 1 M $(NH_4)_2S_2O_8$ aqueous solution is used to etch Cu. The PMMA/graphene is then cleaned with DI water three times. After that, SiO_2/Si substrate is pulled out of the PMMA/graphene, which is floating on DI water, and baked at 80 °C for 12 h. Finally, PMMA is completely removed by acetone.

Preparation of exfoliated graphene sheets

First, graphene sheets are exfoliated on an SiO_2/Si wafer by a traditional standard micromechanical cleavage method. Then the samples are UHV annealed ($<10^{-5}$ Pa) at 650 °C for 30 min and H_2 annealed under 6 Pa at 650 °C for 10 min, in turn. For encapsulated graphene, all the edges of the exfoliated sheets are encapsulated by 100-nm Au, which is patterned by standard micro/nanofabrication (electron beam lithography and thermal evaporation). For defective graphene, defects are introduced into the exfoliated samples by Ar ICP (200 W for 100 s at RT under 6 Pa).

Preparation of defective graphene films

All the initial graphene films are grown by traditional CVD on Cu(111) at 1,050 °C. For weakly defective films, they are treated by Ar plasma at 50 W for 60 s under 6 Pa at RT; for intermediate defective films, Ar plasma at 200 W for 50 s under 6 Pa at RT; for heavily defective films, Ar plasma at 200 W for 300 s under 6 Pa at RT.

ICP post-treatment

For ICP post-treatment with H_2, the open, encapsulated and defective exfoliated graphene sheets, or the CVD graphene films, are all post-treated by H_2 ICP at 2 W for 60 s under 6 Pa at 650 °C ('ICP @ dark'), or 20 W for 60 s ('ICP @ light', as this produces a glow within the plasma). For ICP post-treatment with D_2, all the processes follow the same parameters as above, only replacing H_2 by D_2. For ICP post-treatment with He, the CVD graphene films are treated by ICP at 20 W for 300 s under 6 Pa at different temperatures.

Characterization

Optical measurements. RT Raman scattering is performed using a WiTec/alpha 300R confocal microscope with 532-nm laser under ambient conditions. Variable-temperature Raman scattering is performed with 532-nm laser excitation and a grating spectrometer from Princeton Instruments, with the sample temperature controlled by a closed-cycle Montana Instruments Cryostation C2. The laser power is set below 1 mW to avoid heating. The G and 2D peaks in all the Raman spectra are fitted Lorentzians.

AFM, STM, STS and ARPES measurements

AFM measurements are performed with a Bruker Dimension Fastscan system at tapping mode. The STM of graphene films on Cu(111) are performed under UHV at 77 K. STS differential conductance (dI/dV) point spectra are measured in constant-height mode using standard lock-in techniques at 77 K (f = 773.1 Hz, ΔV_{rms} = 12 mV). The ARPES measurement is performed by a hemispherical analyser (Scienta DA30L) with a He I plasma ultraviolet light source (Fermi Instruments). Samples are cooled down to 15 K during ARPES measurements.

Transport measurements

The electrical transport measurements are performed in a ^{4}He cryostat with a superconducting magnet (Oxford Teslatron 8 T) with source meter (Keithley 6430) and lock-in amplifier (Stanford SR830). All transport data are transformed by tensor inversion[33]. The field effect mobility is extracted from $\mu_{FE} = C^{-1}(d\sigma/dV_g)$ (ref. [34]), where C is the gate capacitance per unit area and calculated as ~1.22×10^{-4} F m^{-2} for a 285-nm SiO_2 layer. Carrier mobility is also calculated by the Hall effect when applying a perpendicular magnetic field. The Hall effect mobility is calculated from $\mu_{Hall} = (L/W) \times (1/R_{xx}) \times (dR_{xy}/dB)$, where B is the applied perpendicular magnetic field, L is the length of the channel, W is the channel width, R_{xy} is the Hall resistance and R_{xx} is the longitudinal resistance. The density of carriers is calculated as $n = C(V_g - V_{Dirac})/e$, where C is the capacitance and e is the electron charge.

Computational details

The projector-augmented wave method[35] as implemented in the Vienna ab initio simulation package (VASP)[36] was used together with the generalized gradient approximation exchange correlation functionals in the Perdew–Burke–Ernzerhof[37] parameterization. First-principles molecular dynamics simulations of graphene/H/Cu(111) systems with different H:C ratio at 298 K, 673 K and 923 K under atmospheric pressure. The supercell used for calculations contains single-layer graphene with 24 carbon atoms and five-layer Cu(111) with 60 copper atoms, together with 6 or 12 hydrogen atoms (H:C = 1:4 or 1:2) inside, and experimental values of lattice parameters of graphene and copper were used[38,39]. The initial surface distance between graphene and Cu(111) topmost layer is set to 6 Å, and there is also a 30-Å vacuum layer above graphene. The minimum distance of hydrogen atoms from the surface in our initial supercell is about 2.38 Å, which guarantees that the hydrogen atoms are not directly bonded. The hydrogen atoms are initially located at the exact centre between graphene and the Cu(111) topmost layer. The plane wave cut-off is set to 600 eV, and Γ-point sampling of the Brillouin zone is employed. The first-principles molecular dynamics calculations are performed in the *NVT* ensemble with a Nosé–Hoover thermostat. A 1-fs step is sufficiently short for the evaluation of the motion equations[40]. Considering the layered nature of the system, the van der Waals correction is used with the density functional theory D3 method[41]. At 650 °C, the supercell with H:C = 1:4 has six hydrogen atoms. The equilibrium of this system results in the distance between the graphene and Cu(111) surfaces decreasing to 3.5 ± 0.5 Å, because of the interlayer van der Waals interaction. As the graphene drops towards the copper, this causes an increase in pressure, making the hydrogen atoms move towards the Cu(111) surface and finally bond to copper atoms on the surface; some of them even diffuse to the interstitial spaces between the topmost copper layers. On increasing the hydrogen density to H:C = 1:2, a proportion of the hydrogen atoms re-bond into hydrogen molecules (H_2) during the simulations, and the movement

of H_2 can eliminate the interlayer interaction. During the simulation, higher temperature causes the surface distance to vary strongly, but the average surface distance between graphene and copper is still larger than the initial one (5–7.5 Å).

Data availability

The data that support the findings of this study are available from the corresponding author on reasonable request.

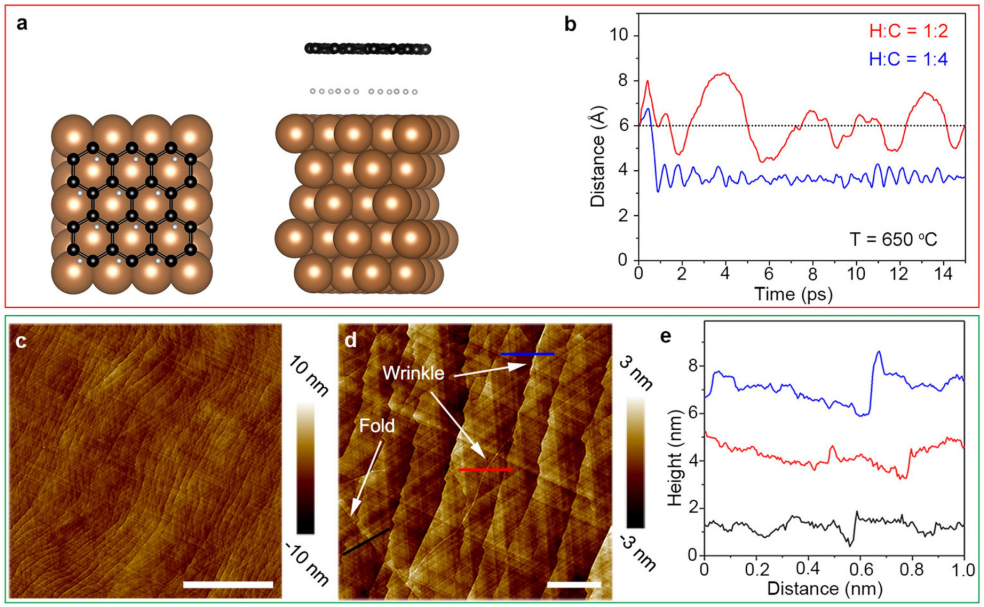
33. Lafont, F. et al. Anomalous dissipation mechanism and Hall quantization limit in polycrystalline graphene grown by chemical vapor deposition. *Phys. Rev. B* **90**, 115422 (2014).
34. Zhang, Y. B., Tan, Y. W., Stormer, H. L. & Kim, P. Experimental observation of the quantum Hall effect and Berry's phase in graphene. *Nature* **438**, 201–204 (2005).
35. Blochl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
36. Kresse, G. & Furthmuller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
37. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett*. **77**, 3865–3868 (1996).
38. Pop, E., Varshney, V. & Roy, A. K. Thermal properties of graphene: fundamentals and applications. *MRS Bull*. **37**, 1273–1281 (2012).
39. Citrin, P. H., Eisenberger, P. & Hewitt, R. C. Adsorption sites and bond lengths of iodine on Cu(111) and Cu(100) from surface extended X-ray-absorption fine-structure. *Phys. Rev. Lett*. **45**, 1948–1951 (1980).
40. Nosé, S. A unified formulation of the constant temperature molecular dynamics methods. *J. Chem. Phys*. **81**, 511–519 (1984).
41. Grimme, S., Antony, J., Ehrlich, S. & Krieg, H. A consistent and accurate ab initio parametrization of density functional dispersion correction (DFT-D) for the 94 elements H–Pu. *J. Chem. Phys*. **132**, 154104 (2010).

Acknowledgements We thank R. Liu for help with electron beam lithography. G.Y., X.H., H.Z., D.W., J.X. and L.G. acknowledge support from the National Key R&D Program of China (grant no. 2018YFA0305800), the NNSF of China (nos. 11674154, 51972163 and 11761131010), the Natural Science Foundation of Jiangsu Province (grant no. BK20190010) and the Fundamental Research Funds for the Central Universities (nos. 020414380094 and 020414380128). D.L. and X. Xi acknowledge support from the National Key R&D Program of China (grant nos. 2017YFA0303201 and 2018YFA0307000) and the NNSF of China (no. 11774151). Y.W. and J.S. acknowledge support from the National Key R&D Program of China (grant no. 2016YFA0300404), the NNSF of China (nos. 11574133, 11974162 and 11834006), the High Performance Computing Center of Collaborative Innovation Center of Advanced Microstructures, the High Performance Computing Center of Nanjing University and "Tianhe-2" in the NSCC-Guangzhou. W.C., X. Xie, J.Z. and Y.Z. acknowledge support from the National Key R&D Program of China (grant no. 2018YFA0306800), the NNSF of China (nos. 11774154 and 11790311) and the Fundamental Research Funds for the Central Universities (no. 020414380110). Q.-Q.Y. and S.-C.L. acknowledge support from the National Key R&D Program of China (grant no. 2014CB921103) and the NNSF of China (nos. 11774149 and 11790311).

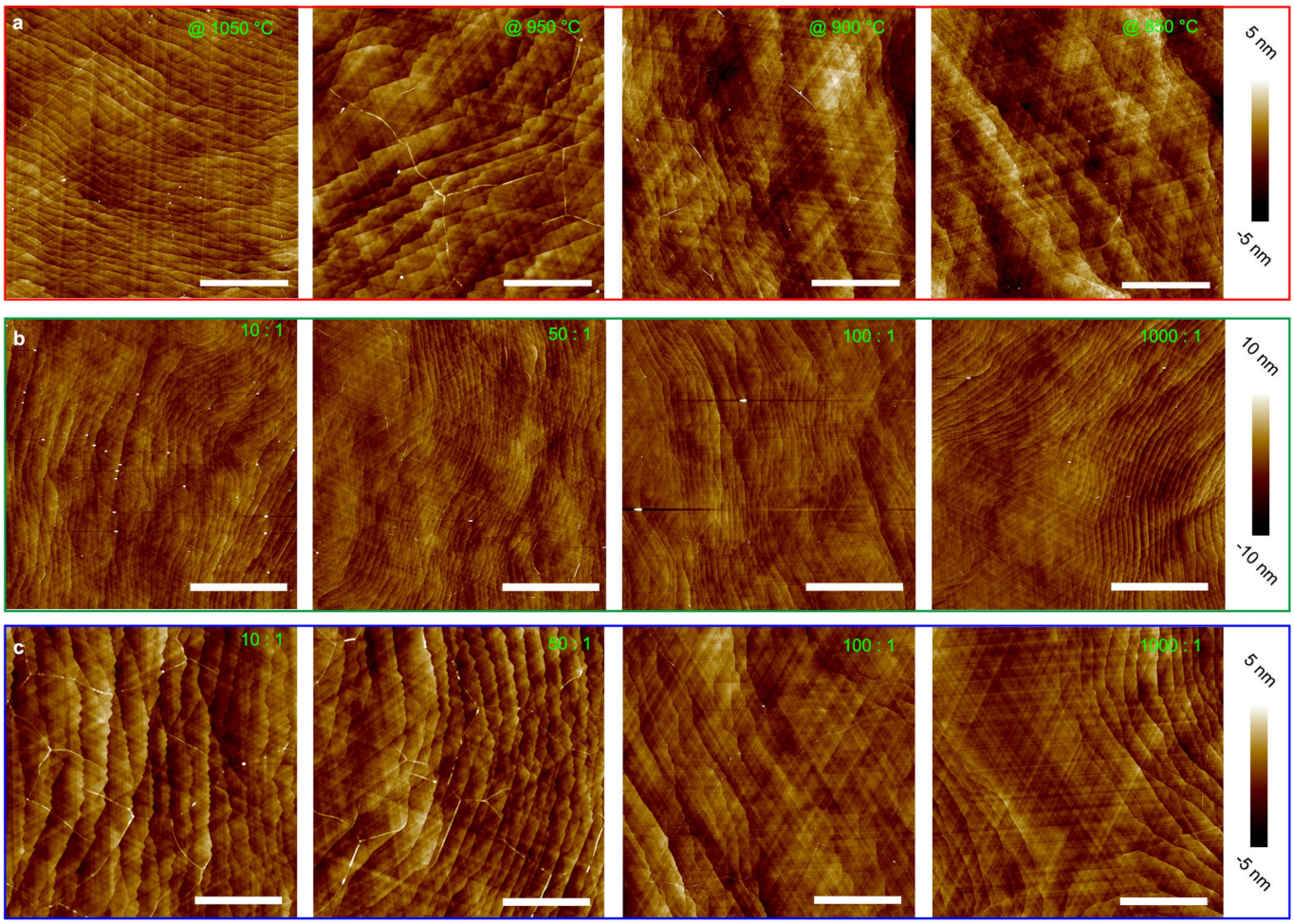
Author contributions L.G. conceived and supervised the project, and designed the experiments. G.Y. performed graphene growth, transfer, AFM and Raman. X.H., H.Z., D.W. and J.X. assisted in the graphene growth, device fabrication, transport and AFM measurements. D.L. and X. Xi performed the variable-temperature Raman measurements. Y.W. and J.S. performed theoretical simulations. W.C., X. Xie, J.Z. and Y.Z. performed ARPES measurements. Q.-Q.Y. and S.-C.L. performed STM and STS measurements. L.G. and G.Y. wrote the manuscript, X. Xi, J.S. and Y.Z. revised it, and all authors commented on it.

Competing interests The authors declare no competing interests.

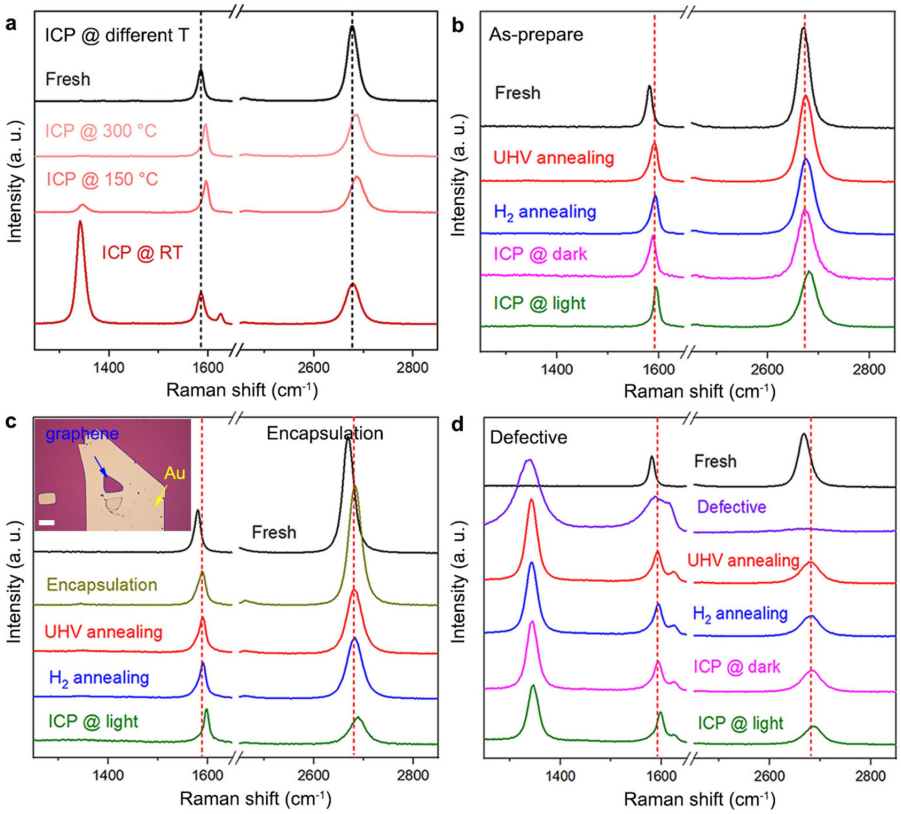
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1870-3.
Correspondence and requests for materials should be addressed to L.G.
Peer review information *Nature* thanks Marcelo Lozada-Hidalgo and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



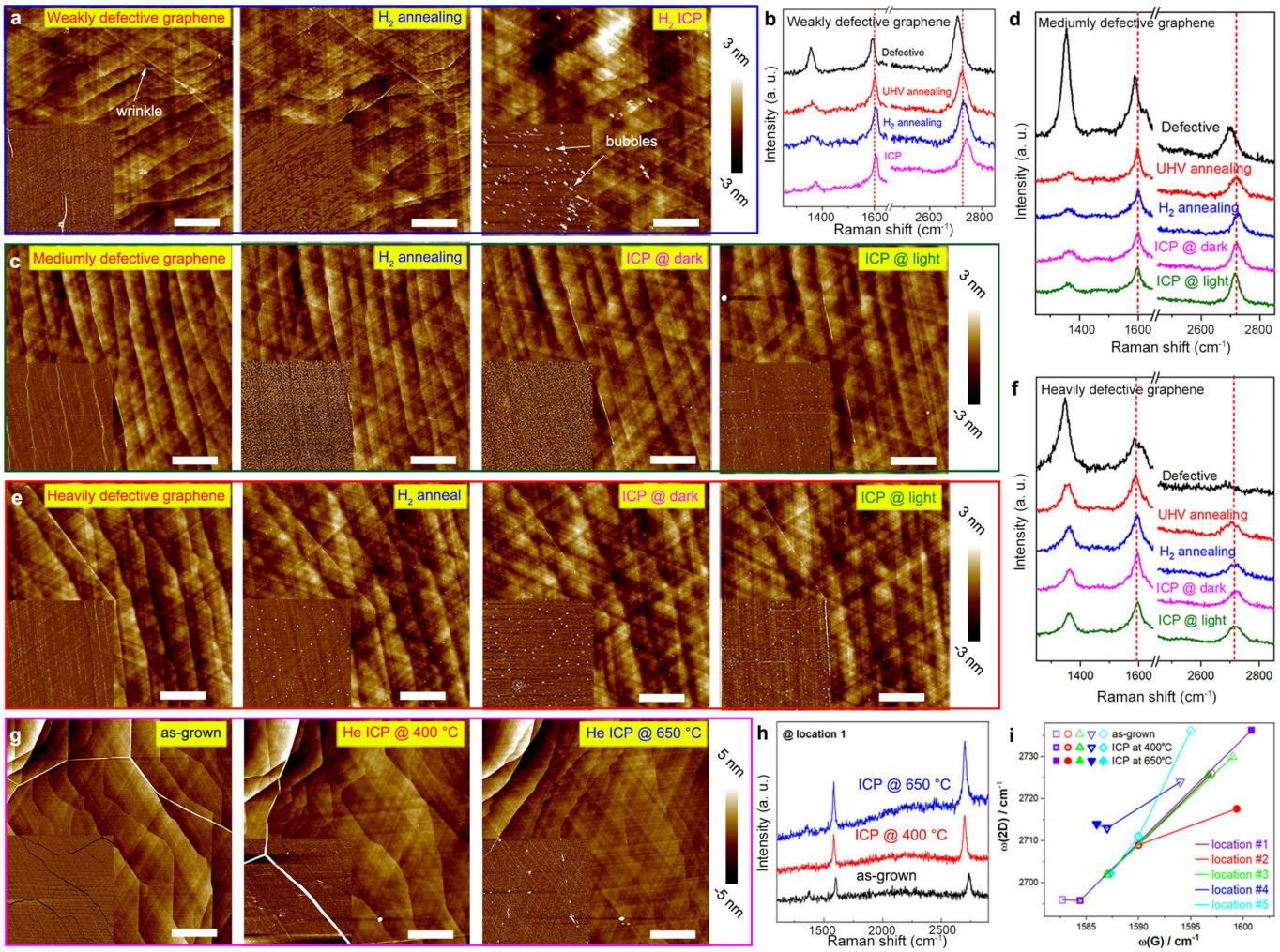
Extended Data Fig. 1 | Theoretical simulation and wrinkle recognition. **a**, Top and profile view of the supercell of graphene, hydrogens and Cu(111). **b**, Theoretical simulation of bonding status for recombined hydrogen with H:C = 1:4 and 1:2 at 650 °C. Some of the hydrogen atoms form into H_2 and increase their surface distance. **c**, AFM image of graphene grown on Cu(111) substrate (region 30 μm × 30 μm). Wrinkles are not apparent. **d**, Enlarged AFM image (5 μm × 5 μm). Wrinkles can be recognized by adjusting the height scale. **e**, Height profiles extracted from **d**. The wrinkles are usually lower than 1 nm. The root-mean-square (RMS) roughness is about 0.756 nm for **c** and 0.538 nm for **d**. Scale bars in **c**, **d** are 10 μm and 1 μm, respectively.



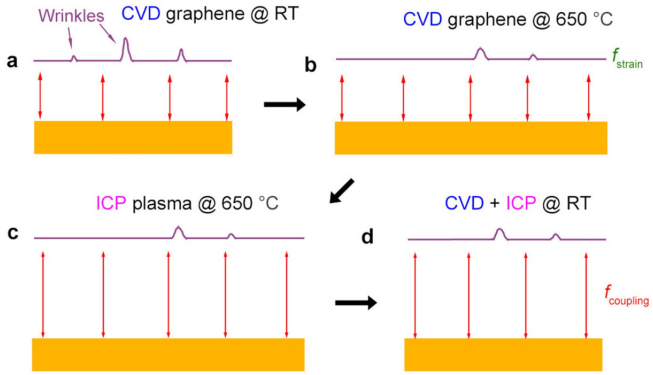
Extended Data Fig. 2 | Wrinkles occur even when growth is at lower temperature or higher H:C ratio. a, AFM images of graphene films grown on Cu(111) at different growth temperatures from 1,050 °C to 850 °C. **b**, Typical AFM images of graphene grown on Cu(111) at 1,000 °C, adjusting the ratio of H_2:CH_4 from 10:1 to 1,000:1. **c**, Enlarged AFM images of **b** with zoom-in height scale. Scale bars in **a**–**c** are 3 μm, 10 μm and 3 μm, respectively.



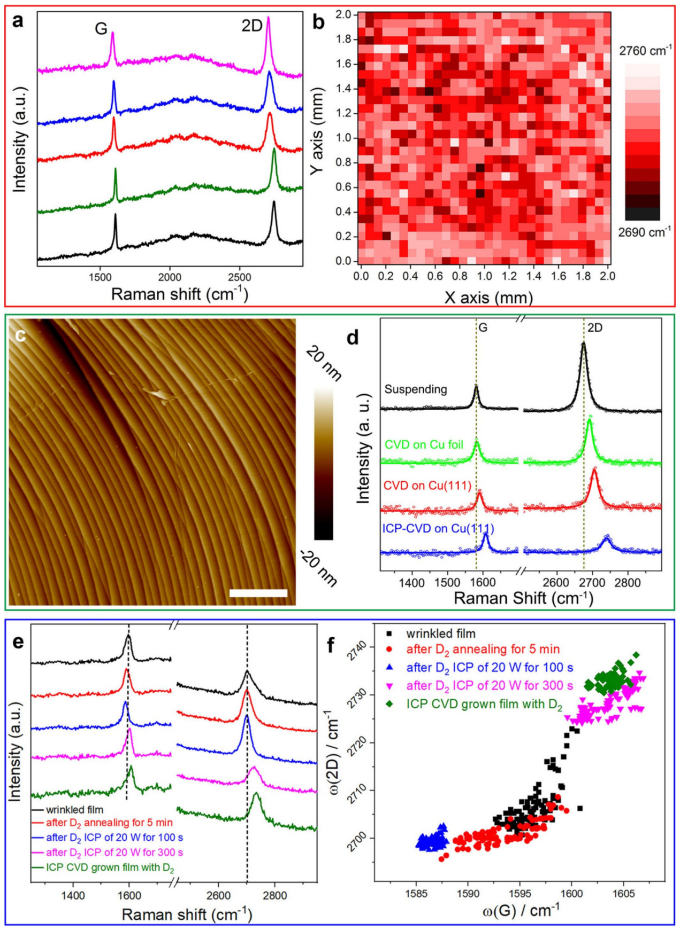
Extended Data Fig. 3 | ICP post-treatment of exfoliated graphene sheets on SiO_2/Si. a, Raman spectra of monolayer graphene after ICP post-treatment at different temperatures *T*. **b**, Intrinsic graphene after post-treatment at different conditions. The Raman spectra are all collected at the same position. **c**, Encapsulated graphene after different treatments. Scale bar, 10 μm. **d**, Defective graphene sheets after different treatments.



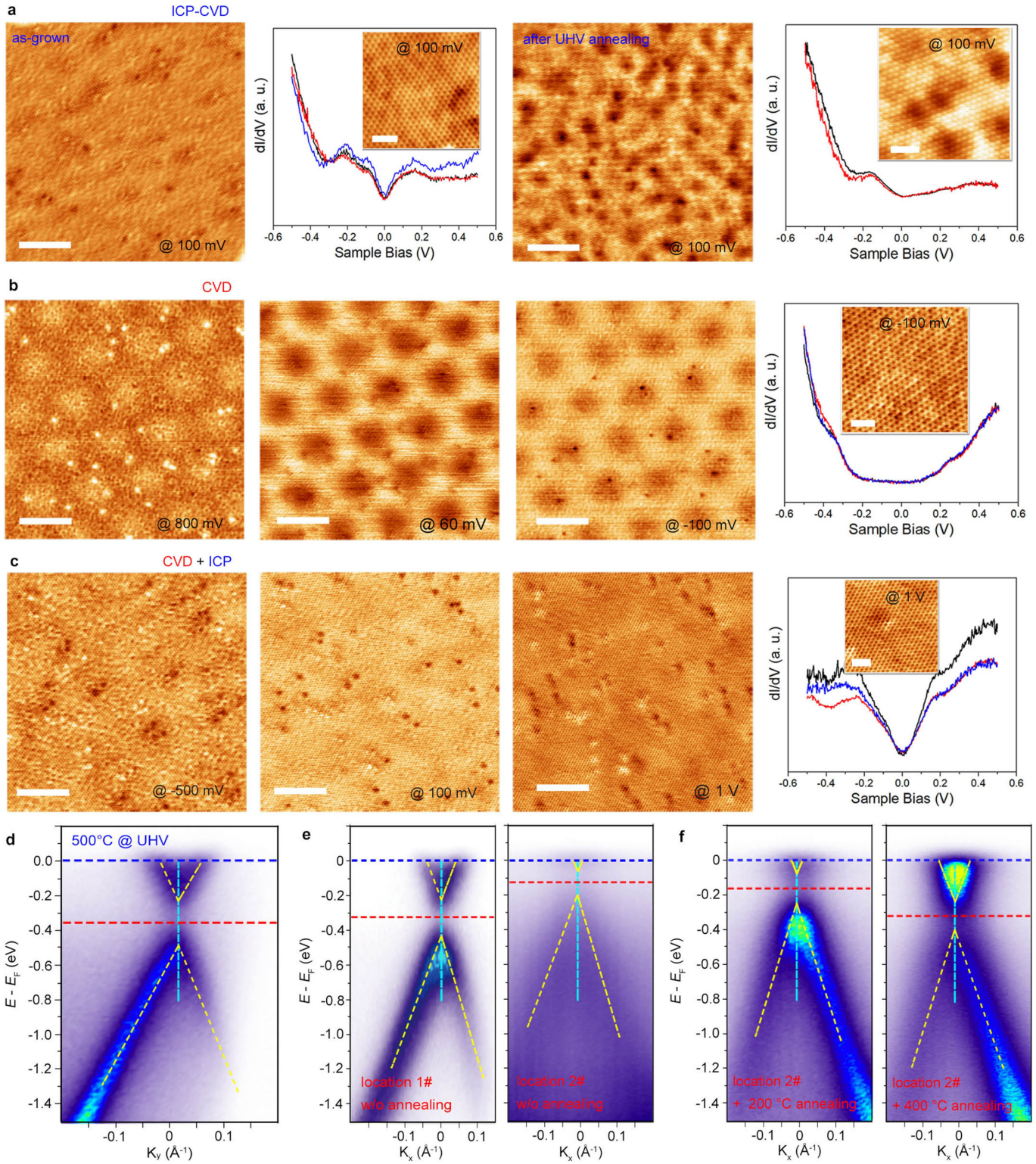
Extended Data Fig. 4 | ICP post-treatment of different graphene films on Cu(111). a, AFM image of weakly defective graphene films after UHV annealing, subsequent H_2 annealing and H_2 ICP with light (see Methods) treatment. Inset, the related AFM phase images. The bubbles apparently occur only after ICP treatment. **b**, Raman spectra of **a** collected at similar position. The Raman frequencies show an obvious shift after H_2 ICP treatment. **c**, AFM image of films with a medium defect level, after UHV annealing, subsequent H_2 annealing, H_2 ICP @ dark treatment and H_2 ICP with light treatment. **d**, Raman spectra of **c** collected at the similar position, showing almost no shift. **e**, AFM image of the heavily defective films after UHV annealing, subsequent H_2 annealing, H_2 ICP with dark treatment and H_2 ICP with light treatment. **f**, Raman spectra of **e** collected at similar position, showing no obvious shift. **g**, AFM image of the as-grown wrinkled graphene films on Cu(111), after He ICP treatment at 400 °C and 650 °C. **h**, Raman spectra collected from the similar position of **g**. **i**, Raman shifts collected from five different positions, showing random shift and no defects formed (no increase in the intensity at ~1,350 cm^{-1} of the D band). All the above results show that the atomic hydrogen, molecular hydrogen, He and He^+ are not helpful in decoupling the defect-free graphene or weakly defective graphene, so only protons (and electrons) in an H_2 plasma will permeate CVD-grown graphene to fully decouple it from the substrates. Scale bars in **a**, **c**, **e** and **g** are all 1 μm.



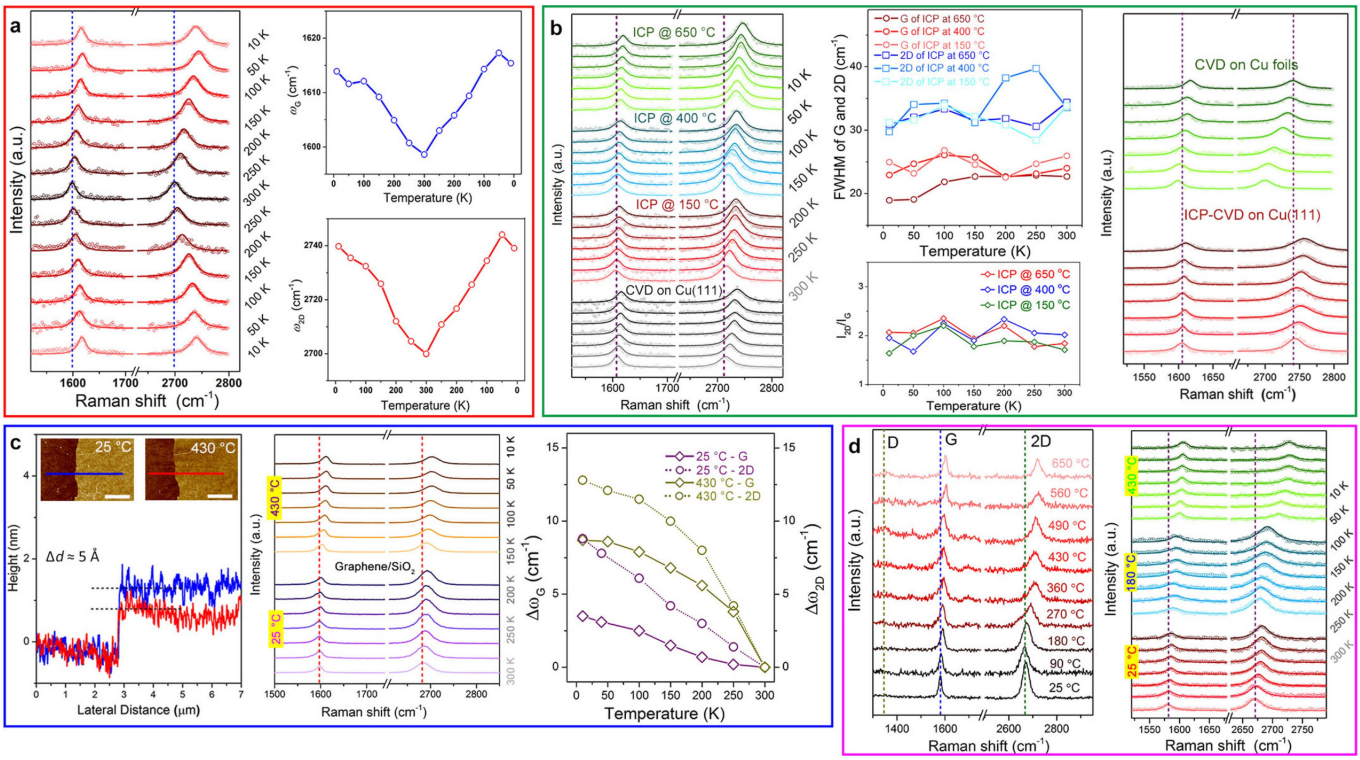
Extended Data Fig. 5 | Proposed mechanism of the irremovable wrinkles by ICP post-treatment. Schematic illustrations of evolution in wrinkles by ICP treatment. **a**, Wrinkles formed to release the high f_{strain}, along with the strong $f_{coupling}$. **b**, Partly released wrinkles when heated to 650 °C, with weak f_{strain} and strong $f_{coupling}$. **c**, Retained wrinkles after ICP treatment at 650 °C, with weak f_{strain} and weak $f_{coupling}$. **d**, Partly released wrinkles when cooling down, with stronger f_{strain} and weak $f_{coupling}$.



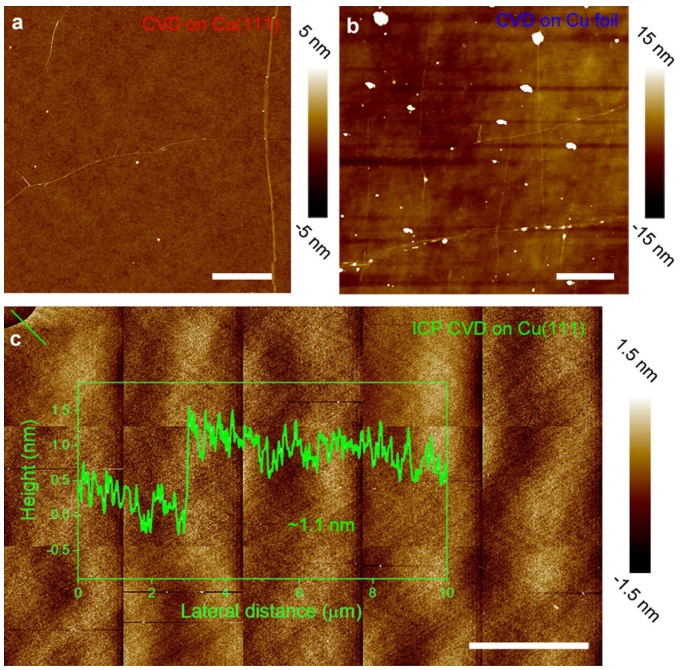
Extended Data Fig. 6 | Inhomogeneous strain in ICP post-treated graphene, comparison of graphene on different Cu substrates and D_2-assisted flattening or growth. a, Typical Raman spectra of wrinkled graphene film after ICP treatment at 650 °C. **b**, Inhomogeneous distribution of ω_{2D} collected from 2 mm × 2 mm region, showing that the strains are inhomogeneous on a large scale. **c**, Typical AFM image of Cu foil fully covered by graphene films. Scale bar, 1 μm. **d**, Typical Raman spectra of graphene films grown on different Cu substrates or suspended on a micro-sized hole. **e**, Raman spectra of wrinkled graphene films on Cu(111) after different treatments and of ultra-flat graphene films grown by ICP–CVD. All the processes here use D_2. **f**, Distribution of Raman frequency collected from multiple points after different treatments.



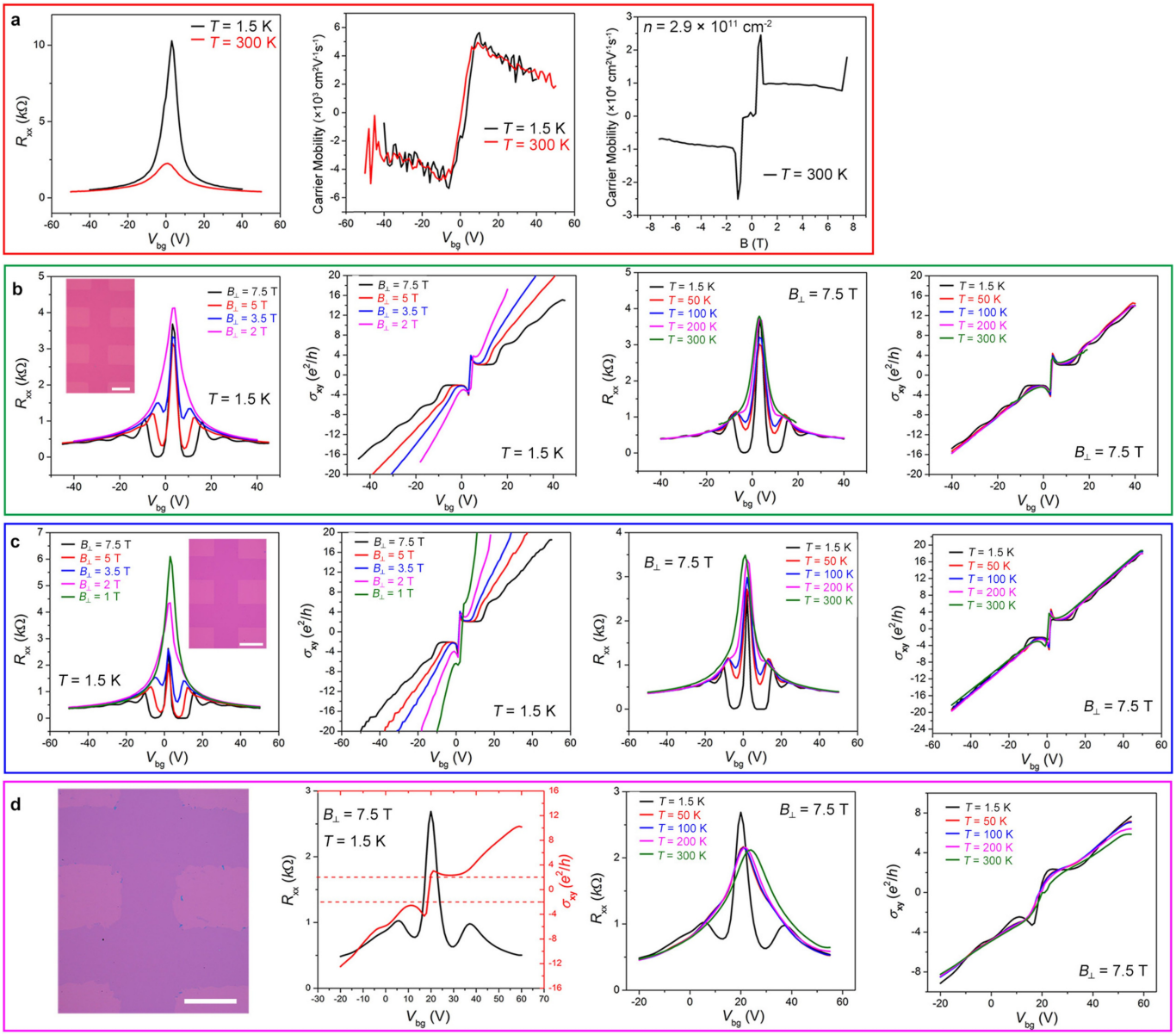
Extended Data Fig. 7 | More STM, STS and ARPES of graphene films on Cu(111). a, Large-scale STM images of as-grown graphene films and after UHV-annealing at 250 °C, grown by ICP–CVD, and their related typical STS spectra. Insets are zoomed-in STM images, showing no periodic moiré pattern and no energy gap. **b**, Large-scale STM images of wrinkled graphene films grown by CVD under different bias voltages, and their related typical STS spectra. Inset is a zoomed-in STM image, showing moiré patterns with period of 5 nm and energy gap of 200–350 meV. **c**, Large-scale STM images of wrinkled graphene films after ICP post-treatment under different bias voltage, and their related typical STS spectra. Inset is zoomed-in STM image, showing no periodic moiré pattern and no energy gap. **d**, ARPES of ICP-CVD graphene films after UHV annealing at 500 °C, indicating the metastable state. **e**, Typical ARPES images extracted from different locations on the ICP post-treated graphene films on Cu(111). The doping is not homogeneous. **f**, The doping level changes gradually after UHV annealing, and the Dirac point will change from −120 meV (without annealing in **e**) to −170 meV (annealing at 200 °C) and finally reaches −320 meV (annealing at 400 °C). Scale bars in **a**–**c** are 4 nm, 6 nm and 4 nm, respectively. All the scale bars in the insets are 1 nm.



Extended Data Fig. 8 | Variable-temperature Raman measurements for grown and transferred graphene on Cu, and transferred graphene on SiO_2/Si. a, Raman spectra with measurement temperature cycled from 10 K to 300 K and from 300 K to 10 K. The Raman frequency of the G band (ω_G) and 2D band (ω_{2D}) is plotted against the variable temperature, demonstrating that the coupling interactions are recovered and this variable-temperature measurement is non-destructive. **b**, Raman frequency shift for graphene films grown on Cu(111) by traditional CVD, before and after ICP treatment, and related FWHM of G and 2D bands and I_G/I_{2D} for the above spectra, showing that FWHM is slightly changed but the I_{2D}/I_G seems almost unchanged; Raman frequency shifts are also shown for as-grown graphene films on Cu(111) by ICP–CVD and on Cu foils by traditional CVD. **c**, Ex situ AFM height profiles for as-transferred and UHV-annealed graphene films. The thickness is decreased by about 5 Å after this UHV annealing. Insets are related AFM height images. In situ variable-temperature Raman measurements are also shown for the transferred and UHV-annealed films on SiO_2/Si wafers, together with related $\Delta\omega_G$ and $\Delta\omega_{2D}$. Scale bars, 2 μm. **d**, Raman spectra of the transferred graphene on Cu(111) under different UHV annealing conditions, and in situ variable-temperature Raman spectra of as-transferred graphene films before (25 °C) and after UHV annealing at 180 °C and 430 °C. The variable temperatures are all from 10 K to 300 K.



Extended Data Fig. 9 | Statistics of residual nanoparticles on different transferred graphene. a–c, AFM images of residual nanoparticles on transferred graphene films after UHV annealing. The films are grown by traditional CVD on Cu(111) (**a**), traditional CVD on Cu foil (**b**) and ICP–CVD on Cu(111) (**c**, 150 μm × 90 μm). Transferred ICP–CVD graphene films appear wrinkle-free and nanoparticle-free during this large-scale measurement. Scale bars in **a**–**c** are 1 μm, 1 μm and 30 μm, respectively.



Extended Data Fig. 10 | Calculation of carrier mobility in the Hall bar devices and the robust QHE in the ultra-flat graphene films. a, Typical transport characteristics of graphene FET with device linewidth of 100 μm measured at 300 K and 1.5 K (left panel); carrier mobilities extracted by field effect $\mu_{FE} \approx 4{,}900\ cm^2 V^{-1} s^{-1}$ and ~$4{,}700\ cm^2 V^{-1} s^{-1}$ for electrons and holes at 300 K, and ~$5{,}600\ cm^2 V^{-1} s^{-1}$ and ~$5{,}300\ cm^2 V^{-1} s^{-1}$ for electrons and holes at 1.5 K (centre panel); and carrier mobilities extracted by Hall effect when applying magnetic field from 0 to 7.5 T at 300 K, showing μ_{Hall} is ~$9{,}800\ cm^2 V^{-1} s^{-1}$ at 300 K with $n \approx 2.9 \times 10^{11}\ cm^{-2}$ (right panel). **b**, Magnetotransport properties of Hall device with linewidth of 20 μm. **c**, Magnetotransport properties of Hall device with linewidth of 100 μm, showing that the first σ_{xy} plateaus appear at $2e^2/h$ for both electrons and holes at RT under 7.5 T. **d**, Magnetotransport properties of Hall device with linewidth of 500 μm, showing that QHE still appears. Insets are optical images of the Hall devices. The device with linewidth 500 μm was fabricated by simple manual operation. Scale bars in **b**–**d** are 20 μm, 100 μm and 500 μm, respectively.

Strain engineering and epitaxial stabilization of halide perovskites

https://doi.org/10.1038/s41586-019-1868-x

Received: 12 April 2019

Accepted: 19 November 2019

Published online: 8 January 2020

Yimu Chen[1,8], Yusheng Lei[1,8], Yuheng Li[1], Yugang Yu[2], Jinze Cai[2], Ming-Hui Chiu[3], Rahul Rao[4], Yue Gu[2], Chunfeng Wang[1], Woojin Choi[5], Hongjie Hu[2], Chonghe Wang[1], Yang Li[1], Jiawei Song[2], Jingxin Zhang[2], Baiyan Qi[2], Muyang Lin[1], Zhuorui Zhang[1], Ahmad E. Islam[4], Benji Maruyama[4], Shadi Dayeh[1,2,5], Lain-Jong Li[3,6], Kesong Yang[1], Yu-Hwa Lo[2,5] & Sheng Xu[1,2,5,7]*

Strain engineering is a powerful tool with which to enhance semiconductor device performance[1,2]. Halide perovskites have shown great promise in device applications owing to their remarkable electronic and optoelectronic properties[3–5]. Although applying strain to halide perovskites has been frequently attempted, including using hydrostatic pressurization[6–8], electrostriction[9], annealing[10–12], van der Waals force[13], thermal expansion mismatch[14], and heat-induced substrate phase transition[15], the controllable and device-compatible strain engineering of halide perovskites by chemical epitaxy remains a challenge, owing to the absence of suitable lattice-mismatched epitaxial substrates. Here we report the strained epitaxial growth of halide perovskite single-crystal thin films on lattice-mismatched halide perovskite substrates. We investigated strain engineering of α-formamidinium lead iodide (α-$FAPbI_3$) using both experimental techniques and theoretical calculations. By tailoring the substrate composition—and therefore its lattice parameter—a compressive strain as high as 2.4 per cent is applied to the epitaxial α-$FAPbI_3$ thin film. We demonstrate that this strain effectively changes the crystal structure, reduces the bandgap and increases the hole mobility of α-$FAPbI_3$. Strained epitaxy is also shown to have a substantial stabilization effect on the α-$FAPbI_3$ phase owing to the synergistic effects of epitaxial stabilization and strain neutralization. As an example, strain engineering is applied to enhance the performance of an α-$FAPbI_3$-based photodetector.

α-$FAPbI_3$ is epitaxially grown on a series of mixed methylammonium lead chloride/bromide ($MAPbCl_xBr_{3-x}$) single crystalline substrates by the inverse temperature growth method[16]. The resulting $MAPbCl_xBr_{3-x}$ substrates, with different compositional ratios and thus lattice parameters, are grown by solutions with different Cl/Br precursor molar ratios (Supplementary Fig. 1)[17]. We note that the strain in the epilayer is determined not only by the lattice mismatch, but also by the relaxation mechanisms. Lattice distortion relaxes the strain, so the region near the heteroepitaxy interface has the highest strain, which gradually drops at regions distant from the interface. The total elastic strain energy increases as the film grows thicker, until it eventually crosses the threshold energy for structural defect generation, and dislocations will form to partially relieve the misfit[18]. A slow growth rate of the epilayer is chosen, as a higher rate will increase the defect concentration in the epilayer. The crystalline quality of the substrates is carefully optimized, as the defects in the substrates can propagate into the epilayer (Extended Data Fig. 1).

Heteroepitaxial growth leads to controllable film thickness, preferential growth sites and orientations, compatible fabrication protocols with existing infrastructures and scalable large-area device applications. Figure 1a shows optical images of a series of $MAPbCl_xBr_{3-x}$ substrates with a layer of epitaxial α-$FAPbI_3$ film on the top. The epilayer has a uniform thickness with a well defined film–substrate interface (Fig. 1b). The film topography can reveal the growth mechanism and sometimes the defects caused by strain relaxation. On the one hand, a sub-100 nm α-$FAPbI_3$ thin film shows a clear interface (Fig. 1b), and a well defined terrain morphology, with a step height close to the size of a α-$FAPbI_3$ unit cell, indicating layer-by-layer growth behaviour of the epitaxial α-$FAPbI_3$ (Extended Data Fig. 2a, b). A 10-μm film, on the other hand, shows non-conformal growth, indicating strain relaxation by dislocation formation (Extended Data Fig. 2c, d).

The crystallographic relationships between the $MAPbCl_xBr_{3-x}$ substrates and the epitaxial α-$FAPbI_3$ thin films are illustrated by high-resolution X-ray diffraction (XRD) (Fig. 1c). In their freestanding form, both α-$FAPbI_3$ and $MAPbCl_xBr_{3-x}$ have a cubic structure[19,20]. The lattice parameters of freestanding α-$FAPbI_3$ and $MAPbCl_xBr_{3-x}$ substrates (both with *Pm*3*m* space group) are calculated to be 6.35 Å (Supplementary Fig. 1) and 5.83–5.95 Å, respectively. The ratio x for each

[1]Department of Nanoengineering, University of California San Diego, La Jolla, CA, USA. [2]Materials Science and Engineering Program, University of California San Diego, La Jolla, CA, USA. [3]Physical Science and Engineering Division, King Abdullah University of Science and Technology, Thuwal, Saudi Arabia. [4]Materials and Manufacturing Directorate, Air Force Research Laboratory, Wright Patterson Air Force Base, Dayton, OH, USA. [5]Department of Electrical and Computer Engineering, University of California San Diego, La Jolla, CA, USA. [6]School of Materials Science and Engineering, University of New South Wales, Sydney, New South Wales, Australia. [7]Department of Bioengineering, University of California San Diego, La Jolla, CA, USA. [8]These authors contributed equally: Yimu Chen, Yusheng Lei. *e-mail: shengxu@ucsd.edu

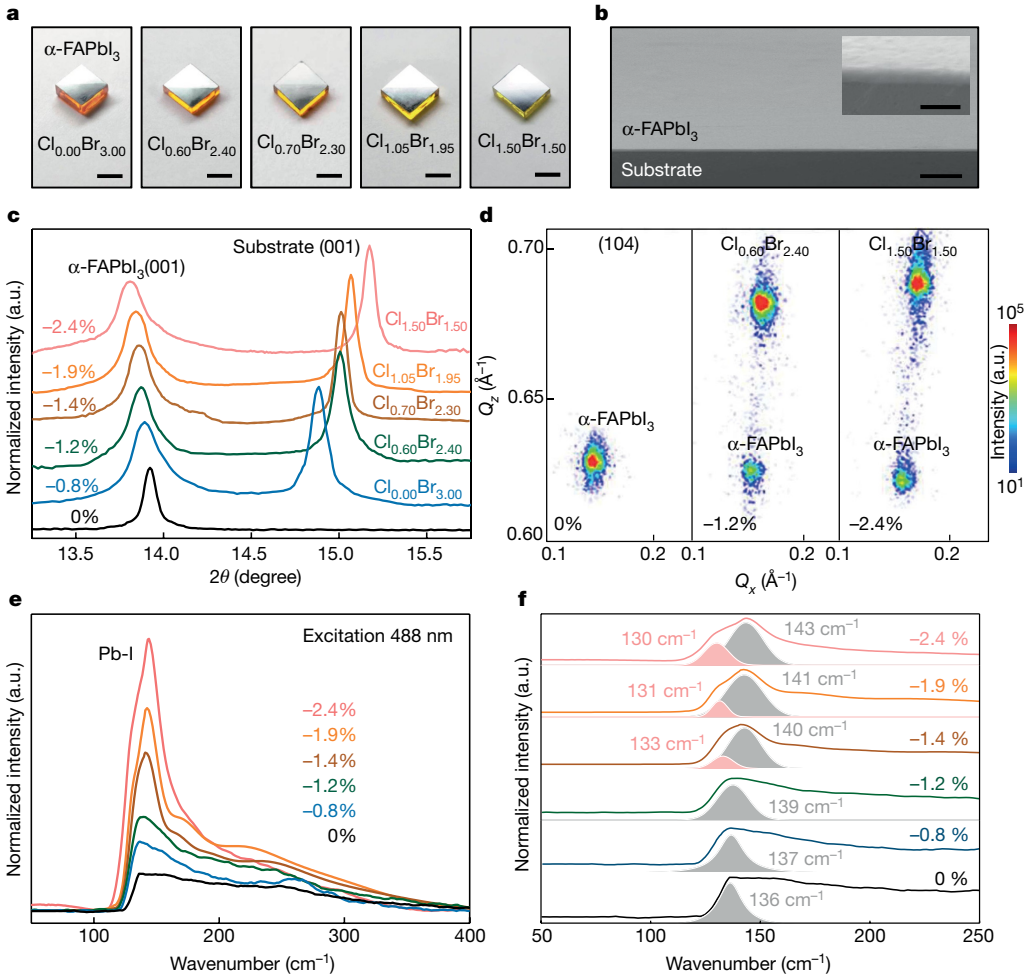


Fig. 1 | Epitaxial α-FAPbI$_3$ thin films and structural characterizations. **a**, Optical images of the as-grown epitaxial α-FAPbI$_3$ thin films. The high transparency of the substrates and the smooth surfaces of the thin films demonstrate their high quality. Scale bars, 4 mm. **b**, A cross-sectional scanning electron microscope (SEM) image of the epitaxial thin film with controlled uniform thickness. Scale bar, 2 μm. Inset, magnified SEM image of the heterostructure showing a well defined interface. Scale bar, 200 nm. **c**, High-resolution XRD $\omega-2\theta$ scan of the (001) peaks of the epitaxial samples on different substrates showing the increasing tetragonality with increasing lattice mismatch. **d**, Reciprocal space mapping with (104) asymmetric reflection of the α-FAPbI$_3$, for different lattice mismatches with the substrate. The results show a decrease in the in-plane lattice parameter as well as an increase in the out-of-plane lattice parameter with larger compressive strain. Q_x and Q_z are the in-plane and out-of-plane reciprocal space coordinates. **e**, Confocal Raman spectra of the epitaxial layer at different strains. We attribute the evolution of the shape and intensity of the peak with strain to the increase in lattice tetragonality under higher strain. We note that the broad peak at approximately 250 cm^{-1} is attributed to the Pb–O bond induced by laser oxidation. **f**, Fitting analysis of the Raman peaks. The peak at 136 cm^{-1} from the strain-free sample (black line) is attributed to the Pb–I bond. With increasing compressive strain, the peak gradually blueshifts as the bond becomes more rigid, and finally splits into a main peak that blueshifts (owing to in-plane bond contraction) and a shoulder peak that redshifts (owing to out-of-plane bond extension). (a.u., arbitrary units).

composition is then calculated to be 0–1.50, according to the Vegard's Law (Supplementary Table 1). As x increases, the MAPbCl$_x$Br$_{3-x}$ (001) peaks shift to a higher 2θ angle, indicating a decrease in the lattice parameters of the substrate and therefore an increase in the lattice mismatch (Fig. 1c and Supplementary Table 2). Meanwhile, the α-FAPbI$_3$ (001) peak shifts to a lower 2θ angle, indicating an increase in the out-of-plane lattice parameter as the in-plane compressive strain increases. When x exceeds 1.50, the strain energy dramatically increases, and the epitaxial growth becomes less thermodynamically favourable. α-FAPbI$_3$ then randomly crystallizes on the substrate (Supplementary Fig. 2). Peak broadening of the epitaxial α-FAPbI$_3$ is therefore induced by the epitaxial strain and the reduction in film thickness, instead of by the strain-induced dislocations or the strain relaxation (Supplementary Fig. 3). Figure 1d shows the reciprocal space mapping of strain-free and strained α-FAPbI$_3$ thin films with different lattice mismatch with the substrate. An increase of tetragonality of the lattice is evident as the compressive strain increases. The corresponding strain levels of the α-FAPbI$_3$ in those three cases are calculated to be 0%, −1.2% and −2.4%, respectively, on the basis of the lattice distortion (where the negative sign denotes compressive strain). The Poisson's ratio is determined to be around 0.3, which is consistent with the reported value[21].

We also studied the structure of α-FAPbI$_3$ at different strains (between 0% and −2.4%, on different substrates) by Raman spectroscopy (Fig. 1e). Control experiments exclude any Raman signals from the substrates (Supplementary Fig. 4). The peak at around 136 cm^{-1} in Fig. 1e, which originated from the stretching of the lead–iodine bond[22], increases in intensity and broadens in width as the strain increases. The cubic structure of the strain-free α-FAPbI$_3$ is less Raman-active, and the detectable signal is usually broad and weak. When in-plane compressive strain increases, the inorganic framework gradually gains tetragonality and produces a stronger Raman signal with a clearly distinguishable shape. Interestingly, at around −1.4% strain, the peak at 136 cm^{-1} starts to split into two: a main peak at about 140 cm^{-1} and a

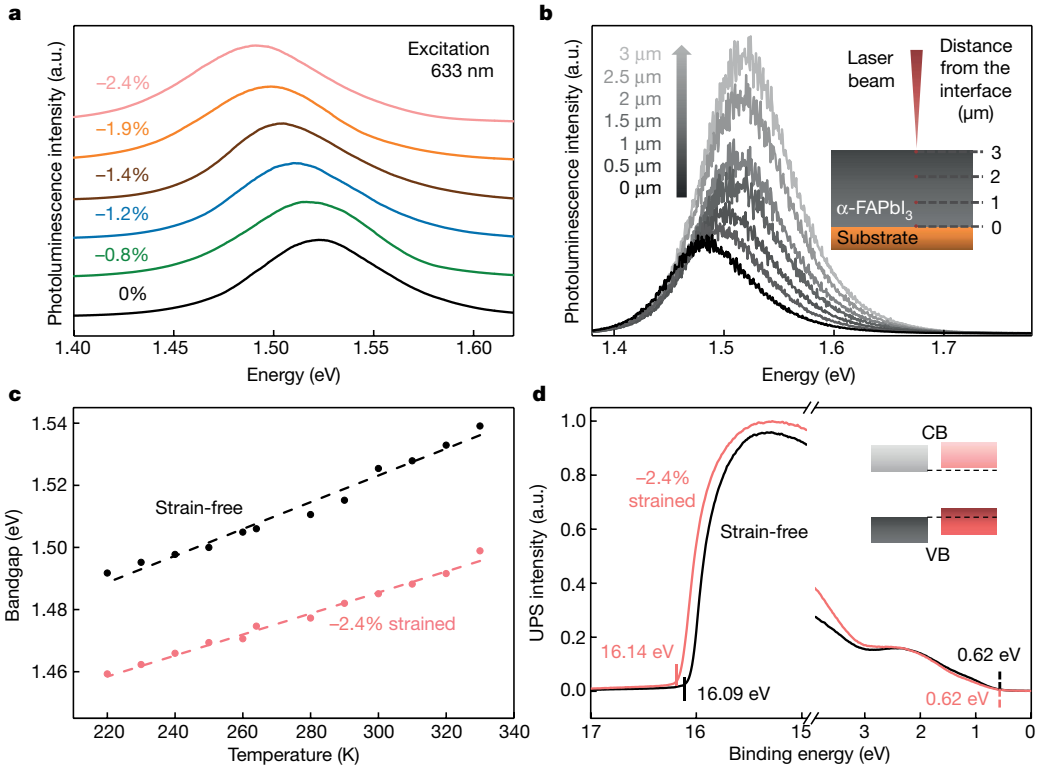


Fig. 2 | Optical properties. a, Photoluminescence spectra of α-$FAPbI_3$ at different strains. The redshift of the photoluminescence peak with increasing strain is due to bandgap reduction under compressive strain, consistent with the first-principles calculations. **b**, Focal-point-dependent confocal photoluminescence spectra of a 3-μm-thick film. When the focal point of the laser (indicated by the red point in the schematic; inset) moves towards the epitaxial interface, the photoluminescence emission peak shifts from about 1.523 eV to about 1.479 eV, owing to the large compressive strain close to the interface. **c**, Temperature-dependent photoluminescence spectra of a −2.4% strained and a strain-free sample. The bandgap of the strain-free sample shows a stronger temperature dependence than the strained sample, indicating that the substrate can reduce the lattice deformation that is caused by the temperature change. **d**, UPS spectra of a −2.4% strained and a strain-free sample. The Fermi level and the VBM of the samples can be extracted from the intersections of the curves with the horizontal axis, marked by the solid and dashed vertical lines, respectively. The results reveal that compressive strain increases the VBM more than it does the CBM, owing to the enhanced interaction of lead 6*s* and iodine 5*p* orbitals under the compressive strain. Inset, the schematic band diagram of the −2.4% strained and strain-free samples. CB, conduction band; VB, valence band.

shoulder at about 133 cm^{-1} (Fig. 1f). When the strain is further increased to −2.4%, these two peaks shift to 143 cm^{-1} and 130 cm^{-1}, respectively. We attribute the blueshift of the main peak to the compression of the in-plane lead–iodine bond, and the redshift of the shoulder peak to the stretching of the out-of-plane lead–iodine bond. This result is also supported by the simulated Raman spectra by first-principles calculations (Supplementary Fig. 4c, d). We also studied the Raman spectra of α-$FAPbI_3$ of various thicknesses on $MAPbCl_{1.50}Br_{1.50}$ (Supplementary Fig. 4f). The results are consistent: a strong, sharp peak is detected from a sub-100-nm film with −2.4% strain, and a weak, broad peak is detected from a film of around 2 μm, where the misfit strain is relaxed near the film surface.

Photoluminescence spectra (Fig. 2a) reveal changes in the bandgap of sub-100-nm epitaxial α-$FAPbI_3$ thin films under different strains (between 0% and −2.4%, on different substrates). The photoluminescence peak of α-$FAPbI_3$ gradually shifts from about 1.523 eV at 0% strain to about 1.488 eV at −2.4% strain, corresponding to a reduction of about 35 meV in the bandgap. We exclude the possible contributions to this photoluminescence redshift from thickness-dependent bandgap[23,24], reabsorption[25] or halide migration[26] (detailed discussions in the Supplementary Information). The bandgap change is consistent with the first-principles calculations and absorption measurements (Extended Data Fig. 3). The photoluminescence peak in Fig. 2a also broadens with increasing strain (Supplementary Fig. 5), which is not due to possible charge transfer between the epitaxial α-$FAPbI_3$ and the substrate (Supplementary Fig. 6). Temperature-dependent photoluminescence studies suggest that the emission peak broadening originates from the reduced crystalline quality and the enhanced carrier–phonon coupling under the strain (Extended Data Fig. 4).

Additionally, we studied confocal photoluminescence spectra at different locations in an α-$FAPbI_3$ film of around 3 μm on a substrate of $MAPbCl_{1.50}Br_{1.50}$ (Fig. 2b). The photoluminescence peak shifts from about 1.479 eV when the laser is focused at the interface where the local strain is high, to about 1.523 eV at 3 μm from the interface where the strain is relaxed. As a control, the photoluminescence redshift in a strain-free sample is less obvious (from about 1.516 eV to about 1.523 eV, Supplementary Fig. 7a), which is attributed to reabsorption[25]. In the strained sample, we exclude elastic relaxation although halide perovskites are much softer than conventional semiconductors[27]. Our finite element analysis simulation results show that the elastic relaxation for a 3-μm-thick α-$FAPbI_3$ thin film is negligible: only around 0.09% (Supplementary Fig. 8). Thickness-dependent in-plane XRD is used to study the critical thickness at which the strain will start to be plastically relaxed (Extended Data Fig. 5). The results show that the critical thickness is much less than the thickness we used in this study and, therefore, the relaxation can be attributed to plastic relaxation by the formation of dislocations. Photoluminescence measurements from samples of different thicknesses show a similar trend (Supplementary Fig. 9), indicating that the strain is relaxed by dislocations when the film grows thicker. Temperature-dependent photoluminescence studies indicate that the bandgap of α-$FAPbI_3$

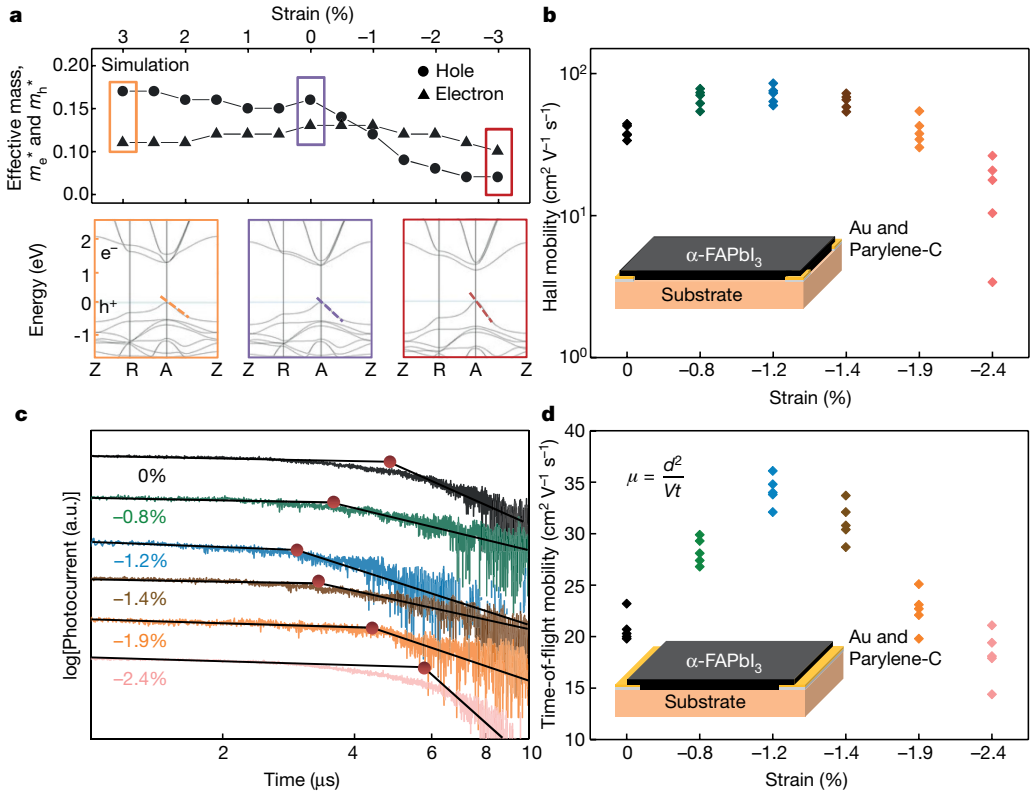


Fig. 3 | Electronic properties. a, Calculated effective masses of the carriers at different strains, and electronic bandstructures under three strain levels (3%, 0% and −3%). The electron effective mass (m_e^*) remains relatively stable with the change in strain, while the hole effective mass (m_h^*) decreases with increasing compressive strain. The dashed lines represent the dispersivity of the valence band; a less dispersive valence bandstructure indicates a smaller hole effective mass. The Z, R and A points are high-symmetry points in the first Brillouin zone of the tetragonal lattice. Bottom panels with coloured borders represent three typical examples with different strains. **b**, Hole mobilities by Hall effect measurements showing that α-$FAPbI_3$ with strain of −1.2% has the highest hole mobility. Coloured symbols correspond to the strain as in **c**. The decrease of the hole mobility with strain higher than −1.2% is attributed to the increase of dislocation density. Number of experiments, $n = 5$ for each strain. Inset, the structure of the measurement setup (gold, yellow; parylene-C, grey), not to scale. **c**, Transient photocurrent curves of the epitaxial α-$FAPbI_3$ under different strains. The transient photocurrent curves are plotted on a log–log scale. The carrier transit time—that is, the inflection point of the photocurrent curve—is marked by a solid red circle. The inflection point indicates the point at which the charge transport carriers switch from the majority to the minority carriers. Lines are guides to the eye. **d**, Plots of calculated carrier mobilities as a function of the strain magnitudes. The inset equation, $\mu = d^2/Vt$, transforms the carrier transit time to the carrier mobility, where μ is the calculated time-of-flight carrier mobility, d is the target region thickness, V is the applied voltage and t is the measured carrier transit time. Number of experiments, $n = 5$ for each strain. Inset, schematic measurement setup. Coloured symbols correspond to the strain as in **c**.

under both 0% and −2.4% strain shows a strong temperature dependence, owing to the soft nature of α-$FAPbI_3$ (Fig. 2c and Extended Data Fig. 4)[7]. The strained-sample bandgap is less temperature-dependent compared to that of the strain-free sample, because the smaller thermal expansion coefficient of the substrate compared to the epitaxial layer introduces a constraint[28] (detailed discussions in the Supplementary Information).

Ultraviolet photoelectron spectroscopy (UPS) reveals the bandstructure evolution of the α-$FAPbI_3$ under strain (see Fig. 2d for 0% and −2.4% strain and Extended Data Fig. 6 for other strains). All samples exhibit p-type behaviour (see Supplementary Information for more details). The Fermi level and the valence-band maximum (VBM) of the samples can be extracted from the UPS data. The results show that strain of −2.4% lifts the VBM upward by about 50 meV compared to the strain-free scenario. Considering the change in the bandgap (about 35 meV, Fig. 2a), the −2.4% strain pushes the conduction-band minimum (CBM) upward by about 15 meV compared to the strain-free scenario. The VBM mainly consists of lead 6*s* and iodine 5*p* orbitals, and the enhanced coupling between these orbitals under compressive strain pushes the VBM upward[29]. The CBM, which consists mostly of nonbonding localized states of Pb *p* orbitals, is less sensitive to the deformation of the PbI_6 octahedrons[7]. Therefore, the in-plane compressive strain increases the VBM more than it does the CBM.

The lattice deformation can alter the electronic bandstructure and therefore also the carrier dynamics. The effective mass of charge carriers can be assessed by the band curvature extracted from first-principles calculations[30]. Figure 3a shows the calculated results of the electron effective mass, m_e^*, and hole effective mass, m_h^* (the top panel) and three typical electronic bandstructures (the bottom panels) under different strains. On the one hand, the E–k dispersion of the conduction band remains relatively unaltered, and m_e^* shows only a slight variation under strain between 3% and −3%. On the other hand, compressive strain can modulate the E–k dispersion of the valence band and considerably reduce m_h^*.

To validate these calculations, Hall effect carrier mobilities of the α-$FAPbI_3$ thin films under strain of between 0% and −2.4% are measured (Fig. 3b). Finite element analysis simulation results show that potential carrier transfer from the substrate to the epitaxial layer is negligible, owing to an insulating layer (Parylene-C) and the energy barrier between the epitaxial layer and the substrate (Supplementary Fig. 10). All samples measured by the Hall effect show a p-type character, which is consistent with the UPS results. Of all strain levels tested, films under −1.2% strain on a $MAPbCl_{0.60}Br_{2.40}$ substrate have the highest hole mobility (Fig. 3b). Further increasing the strain results in a drastic drop in the hole mobility, because of the higher dislocation densities that arise at higher strain levels. We note that the devices for

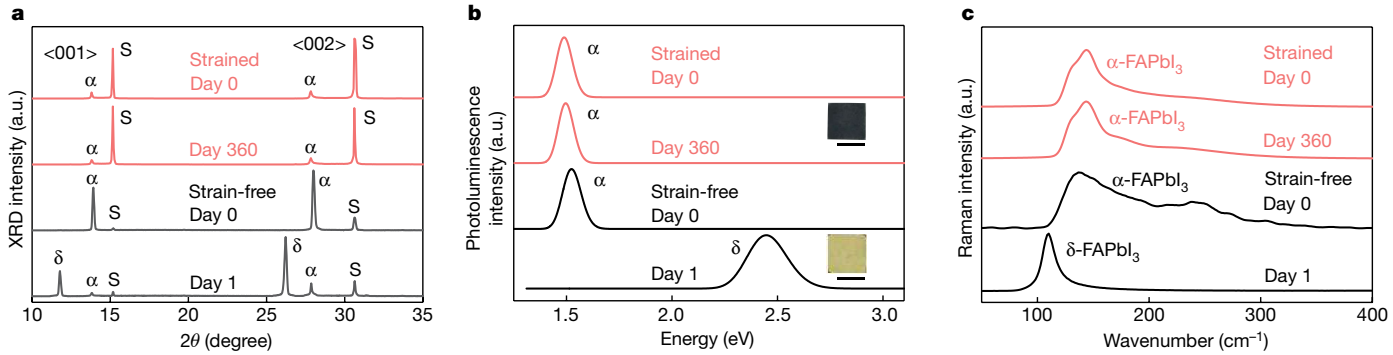


Fig. 4 | Epitaxial stabilization. a, Phase stability comparison of thin (sub-100 nm, −2.4% strained; pink) and thick (about 10 μm, strain-free; black) epitaxial α-$FAPbI_3$ on $MAPbCl_{1.50}Br_{1.50}$ substrates by XRD. α, α-$FAPbI_3$; δ, δ-$FAPbI_3$; S, substrate. The thin, strained sample shows better phase stability (red curves). For the thick, strain-free sample, the (001) peak for α-$FAPbI_3$ at 13.92° is the same as the strain-free sample in Supplementary Fig. 1a, which indicates that the top surface of the thick sample is fully relaxed (day 0, black curve). The X-ray can penetrate about 10–20 μm into the halide perovskites, which explains why the substrate peaks are more intense in the thin sample than in the thick sample. The thick, strain-free sample shows signs of a phase transition to δ-$FAPbI_3$ after 24 h (lower black curve). **b**, Phase stability study by photoluminescence spectroscopy. Re-measurement of the thin, strained sample after 360 d (lower pink curve) shows no obvious photoluminescence peak shift, but does show a slight decrease in peak intensity owing to its natural degradation into PbI_2 (ref. [16]). For the thick, strain-free sample, the photoluminescence spectrum shows an emission peak close to 1.52 eV, similar to that in the strain-free α-$FAPbI_3$ bulk crystal shown in Fig. 2a, indicating a full strain relaxation in the thick sample. Re-measurement after 24 h (lower black curve) shows that the thick film undergoes a transition from the α phase to the δ phase. Insets, optical images of the two samples, showing clear visual clues of the phase stability in the thin, strained sample (black α phase) and the phase transition in the thick, strain-free sample (yellow δ phase) after 24 h. Scale bars, 2 mm. **c**, Phase stability study by Raman spectroscopy. The Raman characteristics of the thin, strained sample show a peak at 143 cm^{-1} with no substantial difference after 360 d; the thick, strain-free sample (peak at 136 cm^{-1}) shows signs of a phase transformation to δ-$FAPbI_3$ after 24 h, as revealed by its signature peak at 108 cm^{-1}.

Hall effect measurements have an epitaxial-layer thickness larger than the critical thickness to ensure sufficient contact area between the halide perovskite and the bottom electrode. Therefore, a high strain level will induce a high concentration of dislocations that degrade the hole mobility.

To validate the Hall mobility, we carried out time-of-flight measurements. The transient photocurrents after single excitation are plotted logarithmically in Fig. 3c. The carrier transit time shows the smallest value of the film under −1.2% strain. The calculated carrier mobility is plotted as a function of the strain applied (Fig. 3d, see the Supplementary Information for calculation details), and shows a similar trend to that given by the Hall effect. We note that the absolute mobility values from the time-of-flight and Hall effect measurements differ, owing to experimental uncertainties in the type and quality of electronic contacts made during the fabrication processes[31]. The space-charge-limited-current method can quantify trap density[32]. Results show that a higher strain level leads to a higher trap density (Extended Data Fig. 7 and Supplementary Fig. 11), which explains the observed decrease in mobility under a higher strain magnitude. Capacitance–frequency (C–ω) spectroscopy is also used to cross-check the trap density (Supplementary Fig. 12), the results of which correspond well with those obtained by the space-charge-limited-current method.

It is widely accepted that α-$FAPbI_3$ crystals are metastable at room temperature and can quickly phase transform to photo-inactive δ-$FAPbI_3$ within approximately 24 h (ref. [16]), owing to its internal lattice strain and low entropy[19,33]. Existing strategies for α-$FAPbI_3$ stabilization, including alloying[26] and surface passivation[34], either widen the bandgap or raise the carrier transport barrier by introducing nonconductive ligands (detailed discussions in the Supplementary Information). However, the epitaxial α-$FAPbI_3$ thin film exhibits long-lasting phase stability at room temperature.

Figure 4a shows XRD results of a sub-100-nm epitaxial α-$FAPbI_3$ thin film that is stable for at least 360 d after growth (red curves in Fig. 4a). In the 10-μm epitaxial thick film (far beyond the threshold thickness at which the strain is fully relaxed), the stabilization effect disappears: after 24 h, XRD peaks from δ-$FAPbI_3$ can be detected (black curves in Fig. 4a). The phase stability of the strained α-$FAPbI_3$ is also verified by photoluminescence (Fig. 4b) and Raman spectroscopy (Fig. 4c). A possible stabilization effect from incorporating bromine or chlorine into the α-$FAPbI_3$ can be excluded, because those foreign ions would stabilize the α-phase regardless of the epilayer thickness. X-ray photoelectron spectroscopy (XPS) measurements showing the absence of bromine and chlorine provide additional evidence that this is not the origin of the stability (Extended Data Fig. 8).

The mechanism of the stable thin α-$FAPbI_3$ can be explained by two reasons. First, the interfacial energy of cubic α-$FAPbI_3$/cubic substrate is much lower than that of hexagonal δ-$FAPbI_3$/cubic substrate, which is the most critical factor for the stabilization effect (Supplementary Fig. 13, Supplementary Table 3, and see Supplementary Information for details). The epitaxial lattice is constrained to the substrate owing to the strong ionic bonds between them and, therefore, the lattice is restricted from the phase transition. Second, the driving force of the α-to-δ phase transition is believed to be the internal tensile strain in the α-$FAPbI_3$ unit cell, which can induce the formation of vacancies and subsequent phase transition[35]. In this study, the epitaxial film is under compressive strain, which neutralizes the effect of the internal tensile strain. Therefore, the synergistic effect of the low-energy coherent epitaxial interface and the neutralizing compressive strain are the key to α-$FAPbI_3$ stabilization. As a control, epitaxial α-$FAPbI_3$ thin film is removed from the substrate (Supplementary Fig. 14); the removed α-$FAPbI_3$ transforms to the δ phase within 24 h.

We demonstrate high-responsivity photodetectors as a use case of the strain engineered α-$FAPbI_3$ thin film. Figure 5a shows the current–voltage (I–V) characteristics of a strain-free device and a device under −1.2% strain. The dark current at −1 V in the strained device is around 15% higher than that in the strain-free one, indicating the higher defect density of the strained device. However, the photocurrent in the strained device increases by approximately 180% compared to the strain-free device. We attribute the photocurrent increase to higher carrier mobility and better alignment of VBM to the Fermi level of the gold electrode under compressive strain (Supplementary Fig. 15).

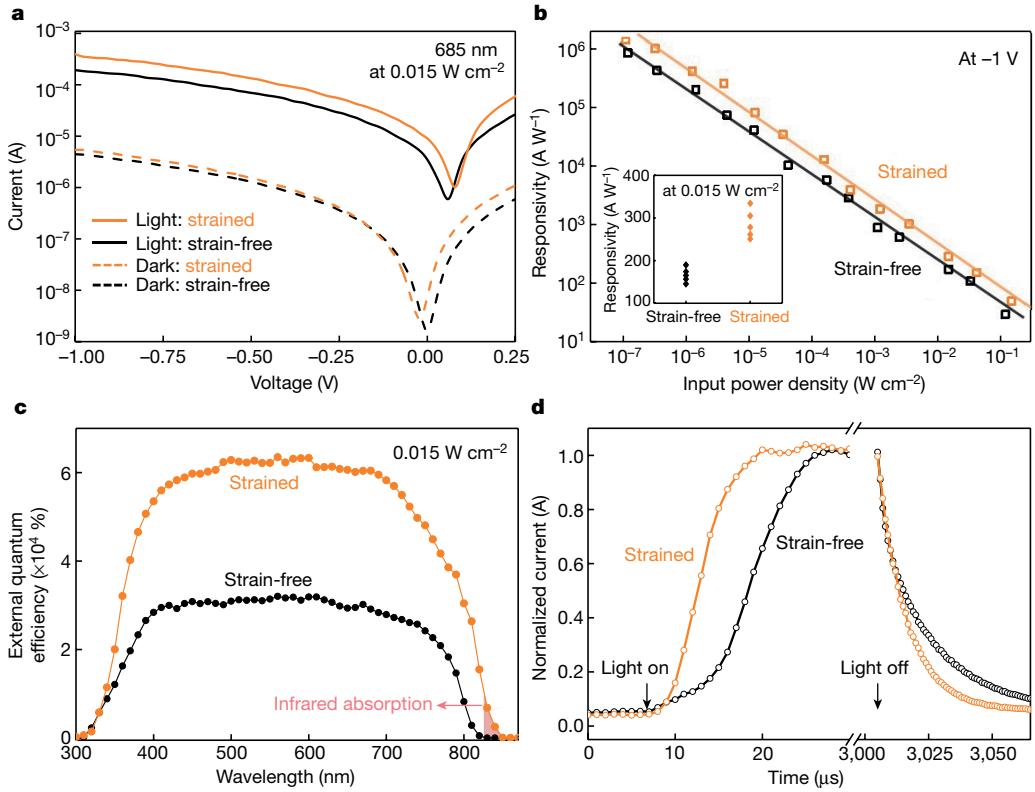


Fig. 5 | Photodetector characterizations of the α-FAPbI$_3$ thin films. **a**, *I*–*V* characteristics of Au/α-FAPbI$_3$/indium tin oxide photoconductor structured photodetectors. The dark current and photocurrent of the −1.2% strained detector are about 15% and 180% higher than those of its strain-free counterpart. Detectors are tested with a 685-nm laser under 0.015 W cm^{-2}. **b**, Comparison of responsivity of the −1.2% strained and strain-free photodetectors. The responsivity of both devices shows an increasing trend with decreasing incident power, as the chances of carrier recombination go down at low illumination intensities[36]. The strained device yields a higher responsivity owing to higher carrier mobility and better band alignment. Inset, the statistical average of the detector performance. Number of experiments, $n = 5$ for each strain value. **c**, External quantum efficiency spectra of the −1.2% strained and strain-free photodetectors showing that the strained photodetector yields a higher external quantum efficiency as well as a broader absorption spectrum (Extended Data Fig. 9d), owing to enhanced carrier mobility and bandgap reduction. **d**, Response times of the photodetectors, with faster rise and fall times for the −1.2% strained (9 μs and 34 μs) than the strain-free (14 μs and 50 μs) device due to the enhanced carrier mobility and transport.

Responsivity of the two photodetectors—defined as the change in photocurrent per unit of illumination intensity—is measured at various illumination intensities (Fig. 5b). The responsivity of the strained device, which reaches a maximum of 1.3×10^6 A W^{-1} at an incident power density of 1.1×10^{-7} W cm^{-2}, is almost twice of that of the strain-free device. This is again attributed to the enhanced carrier mobility and the better band alignment of the strained device. The responsivity of this strained device is, to our knowledge, the highest reported for a α-FAPbI$_3$ device under similar measurement conditions (for example, applied voltage and incident power (Supplementary Table 4)). Similar to the trend in Hall effect carrier mobility, the measured responsivity peaks at −1.2% strain (Extended Data Fig. 9a). Compressive strain also improves the detectivity and the gain of the photodetector (Extended Data Fig. 9b, c). Devices with a diode structure can reduce the dark current, but have a much lower responsivity: on average 500 times lower than that of the photoconductor-type device (Supplementary Fig. 16).

The strained device also shows an enhanced external quantum efficiency over the visible range (Fig. 5c), owing to the enhanced carrier mobility as well as more efficient carrier transport across the gold–perovskite interface. Additionally, after normalizing the spectra, a distinct response in the short-wave infrared region (>810 nm) can be identified for the strained device (Extended Data Fig. 9d), consistent with the photoluminescence measurements showing bandgap reduction under compressive strain. The rise and fall times of the strained device are around 30% shorter than those of the strain-free device, indicating a faster carrier dynamics (Fig. 5d).

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1868-x.

1. Smith, C. S. Piezoresistance effect in germanium and silicon. *Phys. Rev.* **94**, 42–49 (1954).
2. Llordés, A. et al. Nanoscale strain-induced pair suppression as a vortex-pinning mechanism in high-temperature superconductors. *Nat. Mater.* **11**, 329–336 (2012).
3. Yang, W. S. et al. Iodide management in formamidinium-lead-halide-based perovskite layers for efficient solar cells. *Science* **356**, 1376–1379 (2017).
4. Lin, K. et al. Perovskite light-emitting diodes with external quantum efficiency exceeding 20 per cent. *Nature* **562**, 245–248 (2018).
5. Feng, J. et al. Single-crystalline layered metal-halide perovskite nanowires for ultrasensitive photodetectors. *Nat. Electron.* **1**, 404–410 (2018).
6. Lü, X. et al. Enhanced structural stability and photo responsiveness of $CH_3NH_3SnI_3$ perovskite via pressure-induced amorphization and recrystallization. *Adv. Mater.* **28**, 8663–8668 (2016).
7. Liu, G. et al. Pressure-induced bandgap optimization in lead-based perovskites with prolonged carrier lifetime and ambient retainability. *Adv. Funct. Mater.* **27**, 1604208 (2017).
8. Wang, Y. et al. Pressure-induced phase transformation, reversible amorphization, and anomalous visible light response in organolead bromide perovskite. *J. Am. Chem. Soc.* **137**, 11144–11149 (2015).
9. Chen, B. et al. Large electrostrictive response in lead halide perovskites. *Nat. Mater.* **17**, 1164 (2018).
10. Zhao, J. et al. Strained hybrid perovskite thin films and their impact on the intrinsic stability of perovskite solar cells. *Sci. Adv.* **3**, eaao5616 (2017).

11. Zhu, C. et al. Strain engineering in perovskite solar cells and its impacts on carrier dynamics. *Nat. Commun*. **10**, 815 (2019).
12. Steele, J. A. et al. Thermal unequilibrium of strained black $CsPbI_3$ thin films. *Science* **365**, 679–684 (2019).
13. Wang, Y. et al. Nontrivial strength of van der Waals epitaxial interaction in soft perovskites. *Phys. Rev. Mater*. **2**, 076002 (2018).
14. Li, X., et al. Residual nanoscale strain in cesium lead bromide perovskite reduces stability and shifts local luminescence. *Chem. Mater*. **31**, 2778–2785 (2019).
15. Wang, Y. et al. Defect-engineered epitaxial $VO_{2\pm\delta}$ in strain engineering of heterogeneous soft crystals. *Sci. Adv*. **4**, eaar3679 (2018).
16. Han, Q. et al. Single crystal formamidinium lead iodide ($FAPbI_3$): insight into the structural, optical, and electrical properties. *Adv. Mater*. **28**, 2253–2258 (2016).
17. Fang, Y., Dong, Q., Shao, Y., Yuan, Y. & Huang, J. Highly narrowband perovskite single-crystal photodetectors enabled by surface-charge recombination. *Nat. Photon*. **9**, 679–686 (2015).
18. Pohl, U. W. *Epitaxy of Semiconductors: Introduction to Physical Principles* (Springer, 2013).
19. Chen, T. et al. Entropy-driven structural transition and kinetic trapping in formamidinium lead iodide perovskite. *Sci. Adv*. **2**, e1601650 (2016).
20. Maculan, G. et al. $CH_3NH_3PbCl_3$ single crystals: inverse temperature crystallization and visible-blind UV-photodetector. *J. Phys. Chem. Lett*. **6**, 3781–3786 (2015).
21. Rakita, Y., Cohen, S. R., Kedem, N. K., Hodes, G. & Cahen, D. Mechanical properties of $APbX_3$ (A= Cs or CH_3NH_3; X= I or Br) perovskite single crystals. *MRS Commun*. **5**, 623–629 (2015).
22. Steele, J. A. et al. Direct laser writing of δ- to α-phase transformation in formamidinium lead iodide. *ACS Nano* **11**, 8072–8083 (2017).
23. Li, D. et al. Size-dependent phase transition in methylammonium lead iodide perovskite microplate crystals. *Nat. Commun*. **7**, 11330 (2016).
24. Sarmah, S. P. et al. Double charged surface layers in lead halide perovskite crystals. *Nano Lett*. **17**, 2021–2027 (2017).
25. Kanemitsu, Y. Luminescence spectroscopy of lead-halide perovskites: materials properties and application as photovoltaic devices. *J. Mater. Chem. C* **5**, 3427–3437 (2017).
26. Xie, L.-Q. et al. Understanding the cubic phase stabilization and crystallization kinetics in mixed cations and halides perovskite single crystals. *J. Am. Chem. Soc*. **139**, 3320–3323 (2017).
27. Katan, C., Mohite, A. D. & Even, J. Entropy in halide perovskites. *Nat. Mater*. **17**, 377–379 (2018).
28. Ge, C. et al. Ultralow thermal conductivity and ultrahigh thermal expansion of single-crystal organic–inorganic hybrid perovskite $CH_3NH_3PbX_3$ (X = Cl, Br, I). *J. Phys. Chem. C* **122**, 15973–15978 (2018).
29. Yin, W.-J., Yang, J.-H., Kang, J., Yan, Y. & Wei, S.-H. Halide perovskite materials for solar cells: a theoretical review. *J. Mater. Chem. A* **3**, 8926–8942 (2015).
30. Giorgi, G., Fujisawa, J.-I., Segawa, H. & Yamashita, K. Small photocarrier effective masses featuring ambipolar transport in methylammonium lead iodide perovskite: a density functional analysis. *J. Phys. Chem. Lett*. **4**, 4213–4216 (2013).
31. Herz, L. M. Charge-carrier mobilities in metal halide perovskites: fundamental mechanisms and limits. *ACS Energy Lett*. **2**, 1539–1548 (2017).
32. Dong, Q. et al. Electron-hole diffusion lengths > 175 μm in solution-grown $CH_3NH_3PbI_3$ single crystals. *Science* **347**, 967–970 (2015).
33. Zheng, X. et al. Improved phase stability of formamidinium lead triiodide perovskite by strain relaxation. *ACS Energy Lett*. **1**, 1014–1020 (2016).
34. Fu, Y. et al. Stabilization of the metastable lead iodide perovskite phase via surface functionalization. *Nano Lett*. **17**, 4405–4414 (2017).
35. Saidaminov, M. I. et al. Suppression of atomic vacancies via incorporation of isovalent small ions to increase the stability of halide perovskite solar cells in ambient air. *Nat. Energy* **3**, 648–654 (2018).
36. Liu, Y. et al. A 1300 mm_2 ultrahigh-performance digital imaging assembly using high-quality perovskite single crystals. *Adv. Mater*. **30**, 1707314 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Precursor synthesis

Methylammonium bromine (MABr) was synthesized as the precursor for the substrate growth. First, 20 ml methylamine (40% in methanol, Tokyo Chemical Industry) and 21.2 ml hydrobromic acid (48 wt% in water, Sigma Aldrich) were mixed in an ice bath and the temperature was maintained for the reaction to continue for 2 h. The mixture was heated to 80 °C to evaporate the solvent. The precipitate was dissolved in anhydrous ethanol (Sigma Aldrich) at 80 °C and cooled down for recrystallization. The crystals were then centrifuged with diethyl ether and dried at 80 °C overnight.

Crystal growth

Methylammonium lead chloride ($MAPbCl_3$) solution was prepared by mixing 0.6752 g of methylammonium chloride (MACl, 98%, Tokyo Chemical Industry) and 2.781 g lead chloride ($PbCl_2$, 99%, Alfa Aesar) in a mixed solution of 5 ml anhydrous dimethylformamide (99.8%, Aldrich) and 5 ml anhydrous dimethyl sulfoxide (DMSO, 99.8%, Aldrich). Methylammonium lead bromine ($MAPbBr_3$) solution was prepared by mixing 1.120 g MABr and 3.670 g lead bromine ($PbBr_2$, 98%, Acros) in 10 ml dimethylformamide. The $MAPbCl_3$ and $MAPbBr_3$ solutions were mixed with different ratios. The mixed solutions were kept at room temperature to slowly evaporate the solvent, and single crystals were collected to use as substrates. $FAPbI_3$ solutions were prepared by mixing formamidinium iodide (FAI, 99.9%, Greatcell Solar) and lead iodide (PbI_2, 99.99%, Tokyo Chemical Industry) at a molar ratio of 1:1 in anhydrous gamma-butyrolactone (Sigma Aldrich) with different concentrations. Strain-free α-$FAPbI_3$ single crystals were obtained by heating the $FAPbI_3$ solutions to 120 °C.

Epitaxial growth

The substrates were heated to different temperatures, and the preheated $FAPbI_3$ solutions (at 100 °C) were then deposited onto the substrates for epitaxial growth.

Structural and optical characterizations

SEM images were taken with a Zeiss Sigma 500 SEM operated at 3 kV. The $2\theta/\omega$ XRD patterns, the rocking curve (ω scan), and the asymmetrical reciprocal space mapping around the (104) reflection of the substrate were measured by a Rigaku Smartlab diffractometer equipped with a copper $K\alpha_1$ radiation source ($\lambda = 0.15406$ nm) and a germanium (220 × 2) monochromator. The unit cell parameters (a, c) for (104) reflection reciprocal space mapping were converted from (Q_x, Q_z) by $a = 1/Q_x$, $c = 4/Q_z$. Raman and photoluminescence spectra were measured by a Raman spectrometer (Renishaw inVia). Raman peak fitting was done by the Renishaw inVia software. Atomic force microscopy was carried out by a scanning probe microscope (Veeco) in a tapping mode. XPS and UPS were carried out by a Kratos AXIS Supra with an aluminium Kα anode source and a He I (21.22 eV) source, respectively. Measurements were operated under a chamber pressure of 10^{-8} torr. XPS data were calibrated with the $c1s$ peak (284.8 eV). If not otherwise specified, bulk α-$FAPbI_3$ single crystals were used as the strain-free samples for structural and optical characterizations.

Device fabrication

Devices with a vertical structure were fabricated based on a lithography-based method[37]. Parylene-C (50 nm) and gold (50 nm) were sequentially deposited on the substrates, followed by a photolithography process with photoresist AZ-1512. The pattern was composed of an array of 2-μm-diameter circles (exposed) with 1 μm interdistance (covered by photoresist). The gold was chemically etched with wet etchants and the Parylene-C was precisely etched by reactive ion etching. The etched substrates underwent secondary growth in their corresponding growth solutions so that the substrate surface reached the same height as the electrode. Epitaxial growth on the patterned substrate enabled the α-$FAPbI_3$ crystals to initiate from the exposed patterns and gradually merge into a thin film with a controllable thickness. We note that the $MAPbCl_xBr_{3-x}$ substrates were used for the strained devices (heteroepitaxy) and α-$FAPbI_3$ substrates were used for the strain-free devices (homoepitaxy). The top electrodes were then deposited by sputtering (for indium tin oxide, 200 nm). For vertical devices, the area of the top electrode was controlled to be $1 \times 1\ mm^2$ using a shadow mask. For planar devices, Parylene-C (50 nm) and the electrode (gold, 50 nm) were deposited using a shadow mask with designed electrode layouts.

Electrical characterizations

Space-charge-limited-current measurements were carried out by a source meter (Keithley 2400) and a customized probe station in a dark environment. Devices with an Au/Perovskite/Au structure were used. C–ω measurements were carried out by a parameter analyser (B1500, Agilent) in a dark environment. Devices with an Au/perovskite/indium tin oxide structure were used. The thickness of α-$FAPbI_3$ of all devices for space-charge-limited current and C–ω measurements was controlled to be 500 nm. Hall effect measurements were carried out with a Lake Shore Hall measurement system (HM 3000) using the van der Pauw method. We note that the Parylene-C layer prevented direct contact between the substrate and electrodes, eliminating possible carriers extracted from the substrate. The thickness of the α-$FAPbI_3$ for all devices for Hall effect measurement was controlled to be 500 nm. For the time-of-flight measurement, a 685-nm-pulse laser ($10\ mW\ cm^{-2}$) with $<10^{-10}$-s pulse width was used as the light source. Photoresponse was measured with an oscilloscope (MSO6104A Channel Mixed Signal, Agilent). An external bias of 1 V was applied to drive the carriers in the device while a 1-MΩ resistor was connected in series to simulate the open-circuit condition so that the carriers were effectively blocked in the devices[32]. The measurement was carried out in the dark while the bias and the laser power were kept constant. The experiment setup followed the reported time-of-flight measurement of halide perovskite single crystals[32,38–40]. The α-$FAPbI_3$ thickness of all devices for time-of-flight measurements was also controlled to be 500 nm.

Photodetector characterizations

Devices with the structure shown in Supplementary Figs. 15 and 16 were used. A 685-nm laser was used as the light source. The I–V characteristics were collected on a probe station with an Agilent B2912A source meter.

First-principles calculations

First-principles density functional theory calculations were performed using the Vienna ab initio Simulation Package (VASP)[41]. Electron–ion interactions were described using the Projector Augmented Wave pseudopotential[42]. The electron–electron exchange-correlation functional was treated using the Generalized Gradient Approximation parametrized by Perdew, Burke and Ernzerhof[43]. For bandgap calculations, spin–orbit coupling was incorporated owing to the heavy element Pb, and the hybrid functionals within Heyd–Scuseria–Ernzerhof formalism with 25% Hartree–Fock exchange were employed. A cutoff energy of 400 eV for the plane-wave basis set was used. All structures were fully optimized until all components of the residual forces were smaller than $0.01\ eV\ Å^{-1}$. The convergence threshold for self-consistent-field iteration was set at 10^{-5} eV. For optimization of the cubic lattice parameter, a Γ-centred 3 × 3 × 3 k-point mesh was used. A denser k-point mesh of 4 × 4 × 4 was used to obtain accurate energies and electronic structures for strained cells. For optimization and static calculations of the heterostructural models, Γ-centred 4 × 4 × 1 and 5 × 5 × 1 k-point meshes were used, respectively. Raman intensities were calculated by the CASTEP module in Materials Studios[44] with a 3 × 3 × 3 k-point mesh and a 400 eV cutoff energy.

Finite element analysis simulations

Simulation of the current density was done by the multiphysics analysis in COMSOL (version 5.4; www.comsol.com). Simulation of the elastic strain relaxation was done by the ABAQUS[45].

Data availability

The data that support the findings of this study are available from the corresponding authors on reasonable request.

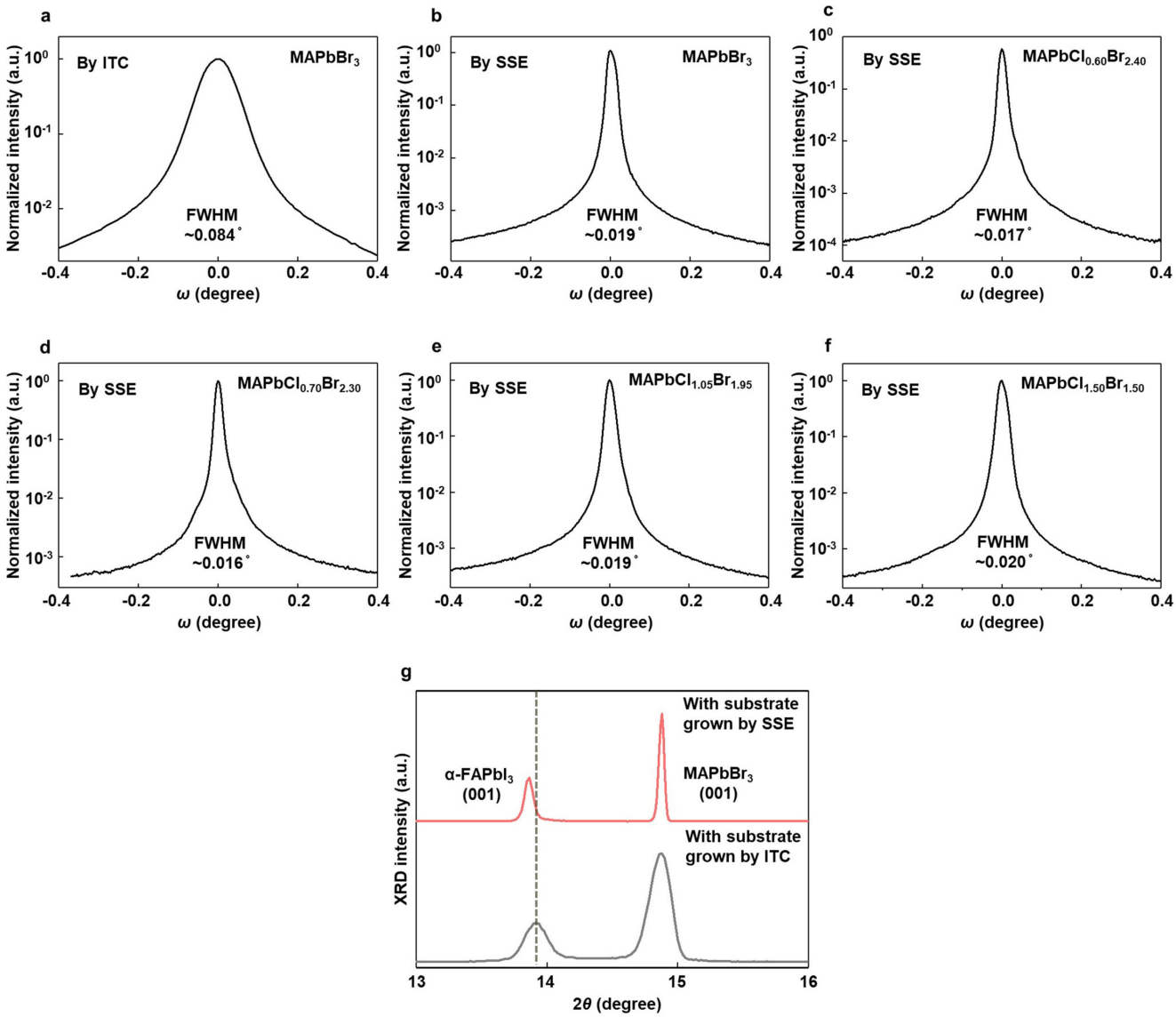
37. Lei, Y. et al. Controlled homoepitaxial growth of hybrid perovskites. *Adv. Mater*. **30**, 1705992 (2018).
38. Shi, D. et al. Low trap-state density and long carrier diffusion in organolead trihalide perovskite single crystals. *Science* **347**, 519–522 (2015).
39. Pospisil, J. et al. Density of bulk trap states of hybrid lead halide perovskite single crystals: temperature modulated space-charge-limited-currents. *Sci. Rep*. **9**, 3332 (2019).
40. Saidaminov, M. I. et al. High-quality bulk hybrid perovskite single crystals within minutes by inverse temperature crystallization. *Nat. Commun*. **6**, 7586 (2015).
41. Kresse, G. & Furthmüller, J. Efficient iterative schemes for ab initio total-energy calculations using a plane-wave basis set. *Phys. Rev. B* **54**, 11169–11186 (1996).
42. Blöchl, P. E. Projector augmented-wave method. *Phys. Rev. B* **50**, 17953–17979 (1994).
43. Perdew, J. P., Burke, K. & Ernzerhof, M. Generalized gradient approximation made simple. *Phys. Rev. Lett*. **77**, 3865–3868 (1996).
44. Clark, S. et al. First principles methods using CASTEP. *Z. Kristallogr*. **220**, 567–570 (2005).
45. Smith, M. *ABAQUS/Standard User's Manual* Version 6.9 (Dassault Systèmes Simulia Corp, 2009).

Acknowledgements We thank T. N. Ng and Z. Wu for guidance on the transient photocurrent measurement; P. Liu and S. Yu for sharing the Rikagu Smartlab diffractometer; D. P. Fenning and X. Li for discussions; Q. Lin for guidance on the reciprocal space mapping measurements; S. Wang for analysis and discussions of the UPS; Y. Zeng for training on the Renishaw inVia Raman spectrometer; Y. Li, Y. Yin and M. Chen for guidance on the finite element analysis simulations; and S. Xiang for constructive feedback on manuscript preparation. This work was supported by the startup fund by the University of California San Diego. The microfabrication involved in this work was performed at the San Diego Nanotechnology Infrastructure (SDNI) of UCSD, a member of the National Nanotechnology Coordinated Infrastructure, which was supported by the the National Science Foundation (grant number ECCS-1542148). K.Y. acknowledges the National Science Foundation under award number ACI-1550404 and computational resources from Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by National Science Foundation grant number ACI-1548562.

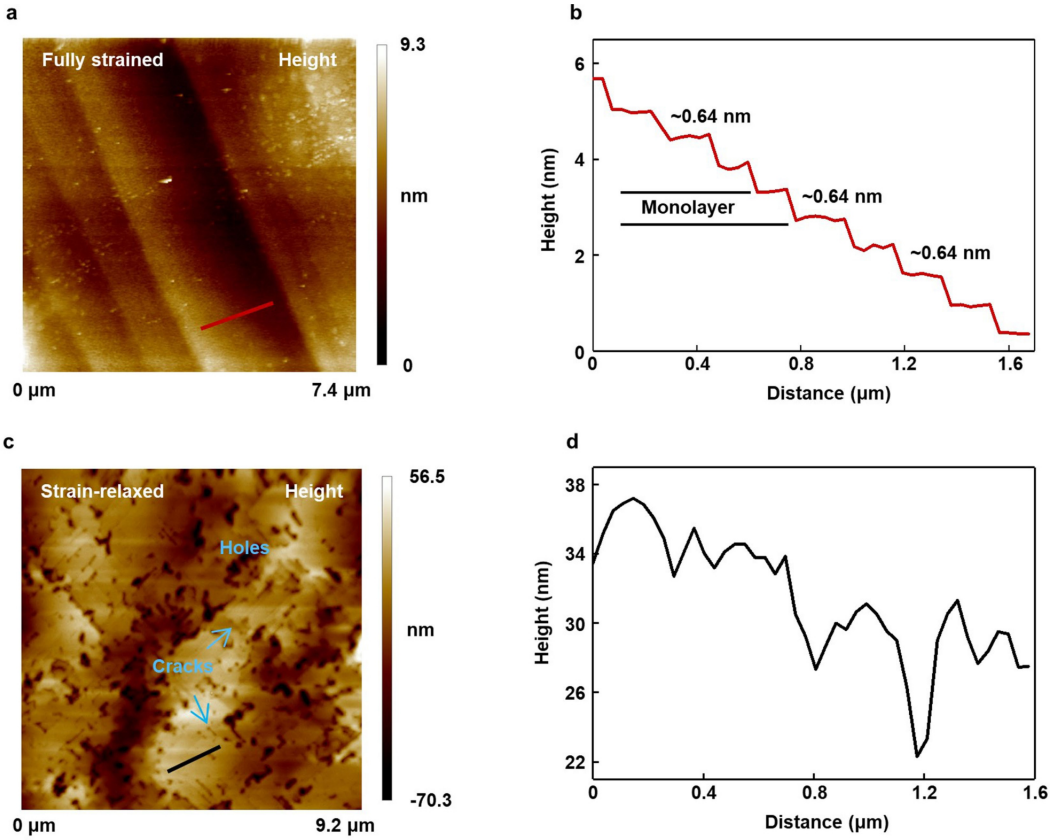
Author contributions S.X. and Y.C. conceived the idea. Y.C. and Y. Lei prepared the samples. Y.C. and Y. Lei took the optical and SEM images. Y.C., J.S. and M.-H.C. carried out the XRD, Raman and photoluminescence spectroscopy characterizations. R.R. and A.E.I. contributed to the temperature-dependent photoluminescence characterizations. Y. Li and J.S. carried out the density functional theory calculations. Y.C. and W.C. carried out the finite element analysis simulations. Y.C., Y. Lei, Y.G., C.W. and J.C. contributed to the device fabrication. Y.C., Y.Y. and W.C. carried out the mobility measurements. Y.C. carried out the trap density measurements. Y.C. and Y.Y. carried out the photodetectors characterizations. All authors contributed to analysing the data and commenting on the manuscript.

Competing interests The authors declare no competing interests.

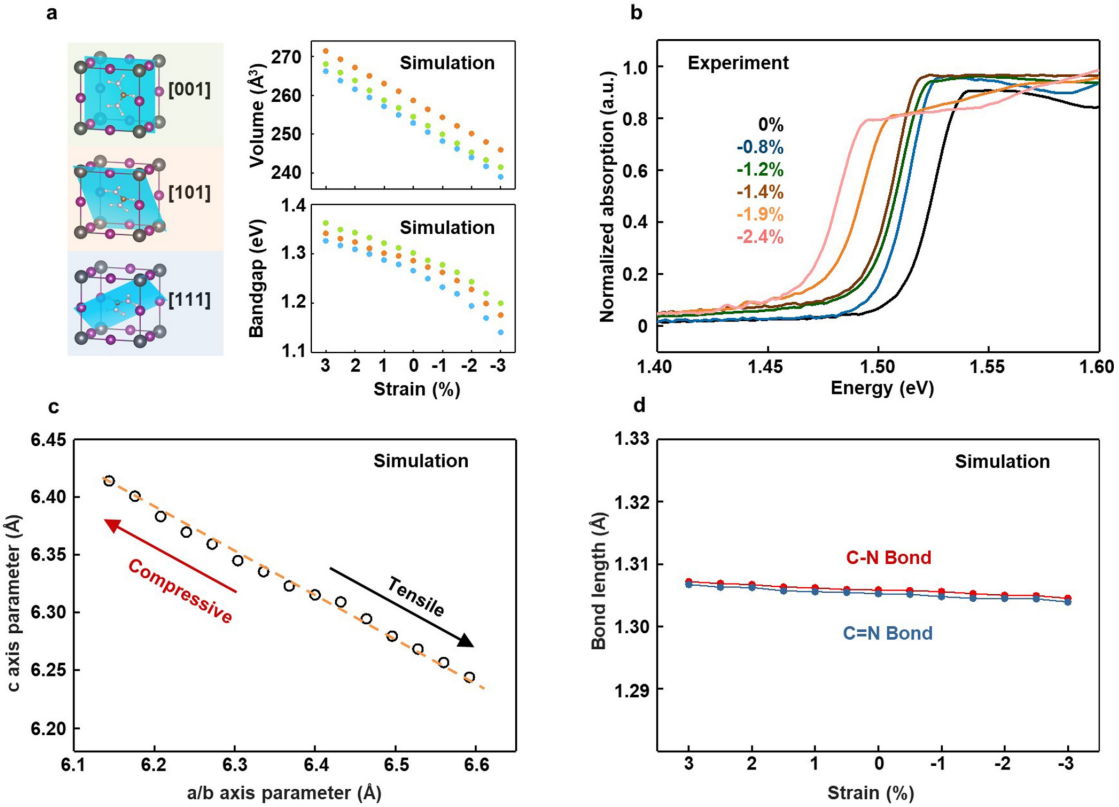
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1868-x.
Correspondence and requests for materials should be addressed to S.X.
Peer review information *Nature* thanks Jian Shi, Lijun Zhang and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



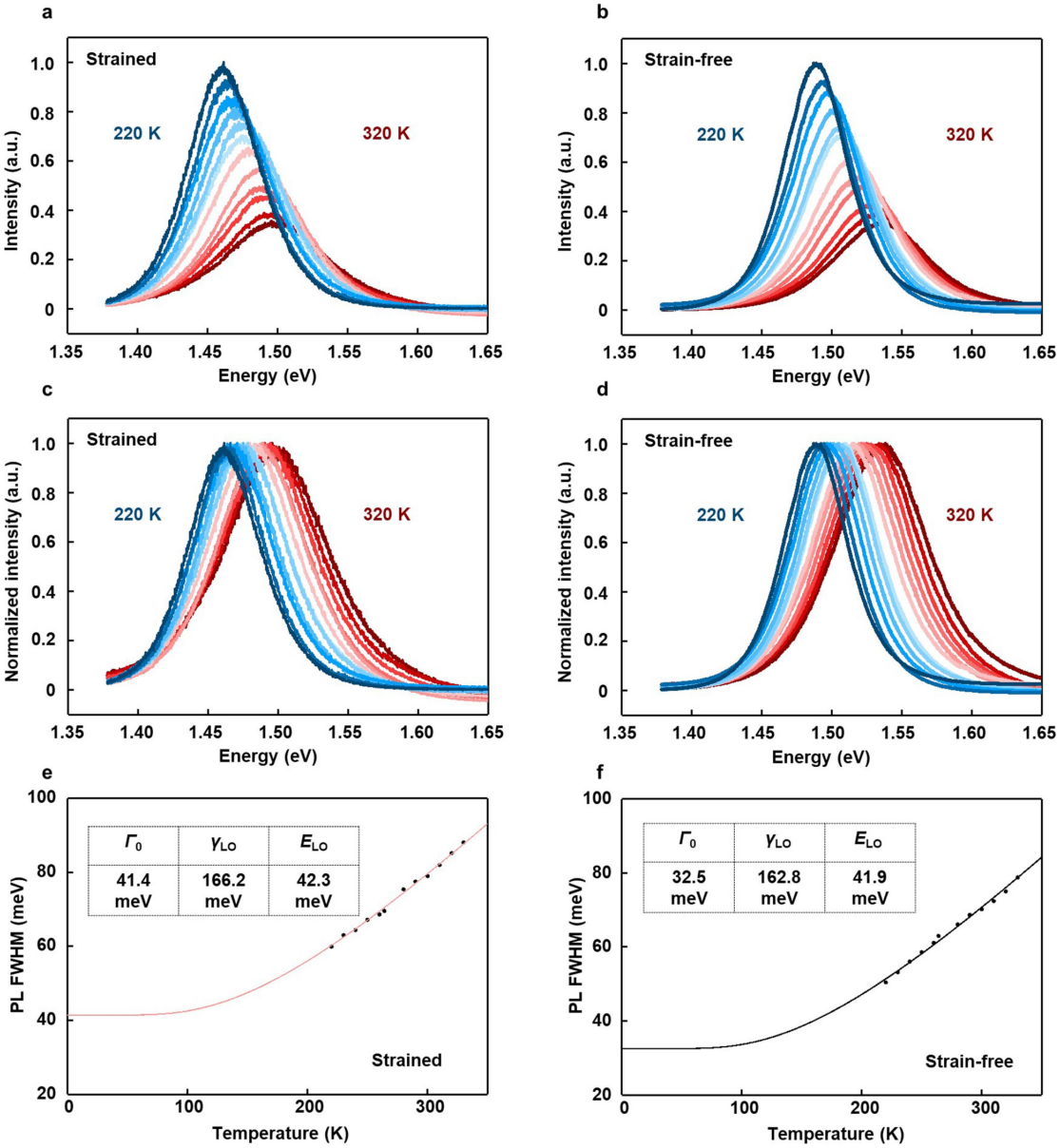
Extended Data Fig. 1 | Characterization of substrate quality with different growth methods and its impact on the epitaxial strain. a–**f**, Rocking curve measurements of substrates grown by the inverse temperature crystallization (ITC) and slow solvent evaporation (SSE) methods. Lower full-width at half-maximum (FWHM) values by the SSE indicate better crystal quality. **g**, XRD patterns of strained α-$FAPbI_3$ on a substrate with higher crystal quality (red curve) and relaxed α-$FAPbI_3$ on a substrate with lower crystal quality (grey curve). Dislocations in the substrates can propagate into and relax the strain in the epitaxial α-$FAPbI_3$. The vertical dash line labels the (001) peak position of strain-free α-$FAPbI_3$. The peak position from the strain-relaxed $FAPbI_3$ (grey curve) shifts back to that of strain-free α-$FAPbI_3$.



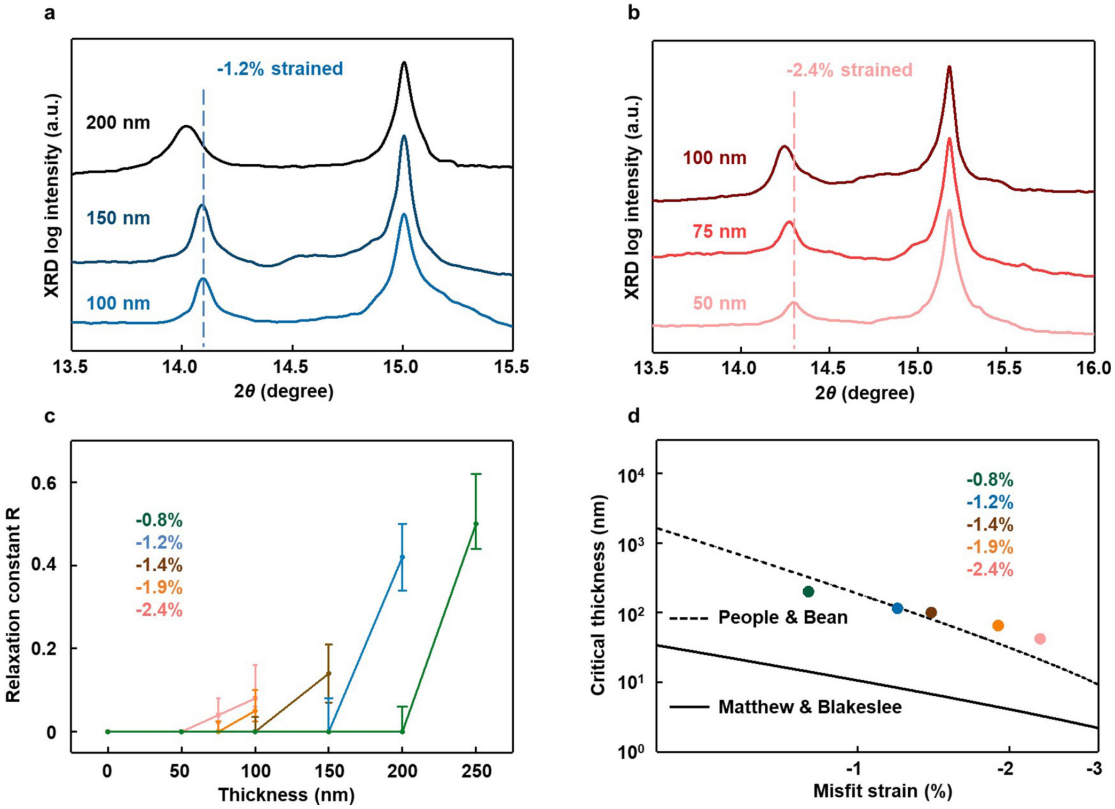
Extended Data Fig. 2 | Atomic force microscopy morphology characterization of strained and strain-relaxed epitaxial α-$FAPbI_3$ films. **a**, **b**, A topography image (**a**) and the corresponding height scanning curve (**b**) of the red line in **a** of a strained epitaxial α-$FAPbI_3$ thin film. **c**, **d**, A topography image (**c**) and the corresponding height scanning curve (**d**) of the black line in **c** of a strain-relaxed epitaxial α-$FAPbI_3$ thick film. Results show that the strained thin film adopts a layer-by-layer growth model. No cracks or holes can be detected. As the film thickness increases, the total strain energy builds up and generates dislocations that propagate throughout the film and relax the strain, leading to the formation of cracks and holes. These cracks and holes are typically regarded as a signature of strain relaxation.



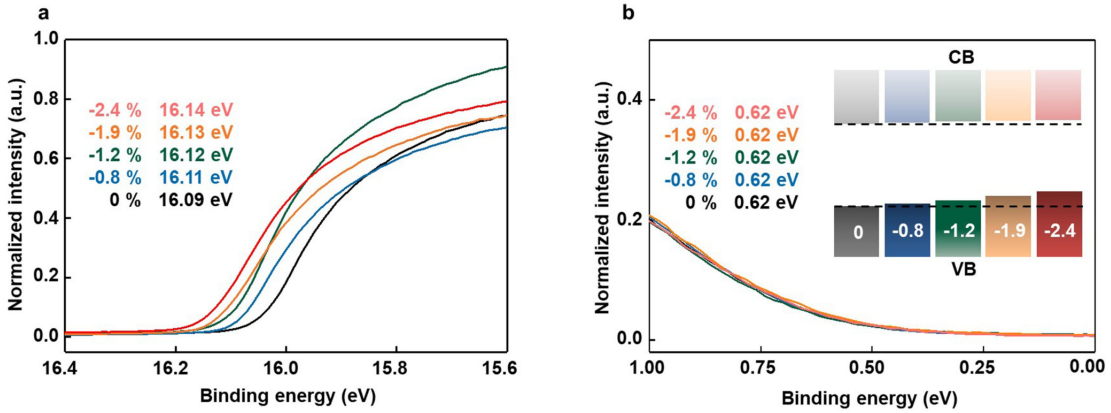
Extended Data Fig. 3 | First-principles calculations of the strained α-$FAPbI_3$ unit cell and experimental absorption spectra of the strained α-$FAPbI_3$ under different strains. a, Evolution of lattice volume and bandgap as a function of strain for three α-$FAPbI_3$ lattices with different FA^+ organic cation orientations. For the bandgap calculations, spin–orbit coupling is incorporated owing to the heavy element Pb, and the hybrid functionals within Heyd–Scuseria–Ernzerhof formalism with 25% Hartree–Fock exchange are employed. **b**, The absorption spectra of the strained α-$FAPbI_3$ thin films. The absorption onset redshifts will the increasing strain, which agrees with the photoluminescence spectra and prove that the strain can alter the bandgap of the α-$FAPbI_3$. **c**, The *c* axis length of the unit cell when biaxially straining the *a*/*b* axes. The slope of the fitted line shows a Poisson's ratio of about 0.3. **d**, C–N and C=N bond lengths at different strain levels. Simulation results show that the deformation of the FA^+ skeleton is very small under the applied biaxial strain.



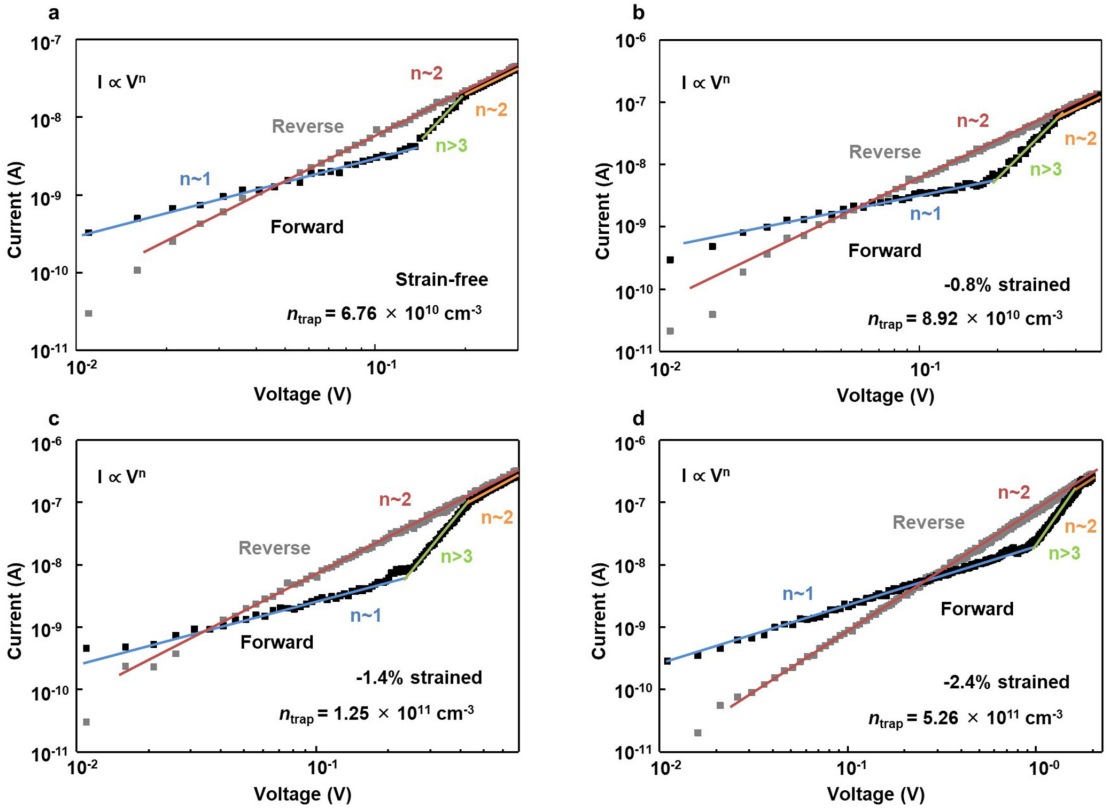
Extended Data Fig. 4 | Temperature-dependent photoluminescence measurement. a, **b**, Temperature-dependent photoluminescence of strained (**a**) and strain-free (**b**) α-$FAPbI_3$ before normalization. **c**, **d**, Temperature-dependent photoluminescence of strained (**c**) and strain-free (**d**) α-$FAPbI_3$ after normalization. Both samples exhibited uniform bandgap narrowing and FWHM narrowing with decreasing the temperature. **e**, **f**, Temperature-dependent photoluminescence (PL) FWHM of strained (**e**) α-$FAPbI_3$ and strain-free (**f**) α-$FAPbI_3$ with fitting. Results show that the strained α-$FAPbI_3$ has a higher Γ_0, γ_{LO} and E_{LO} than that of strain-free α-$FAPbI_3$ owing to the strain-induced crystalline quality reduction and the strain-enhanced carrier-phonon scattering. Γ_0, temperature-independent emission linewidth term associated with the structural disorder scattering. γ_{LO}, charged-carrier-optical-phonon coupling constant. E_{LO}, optical phonon energy.



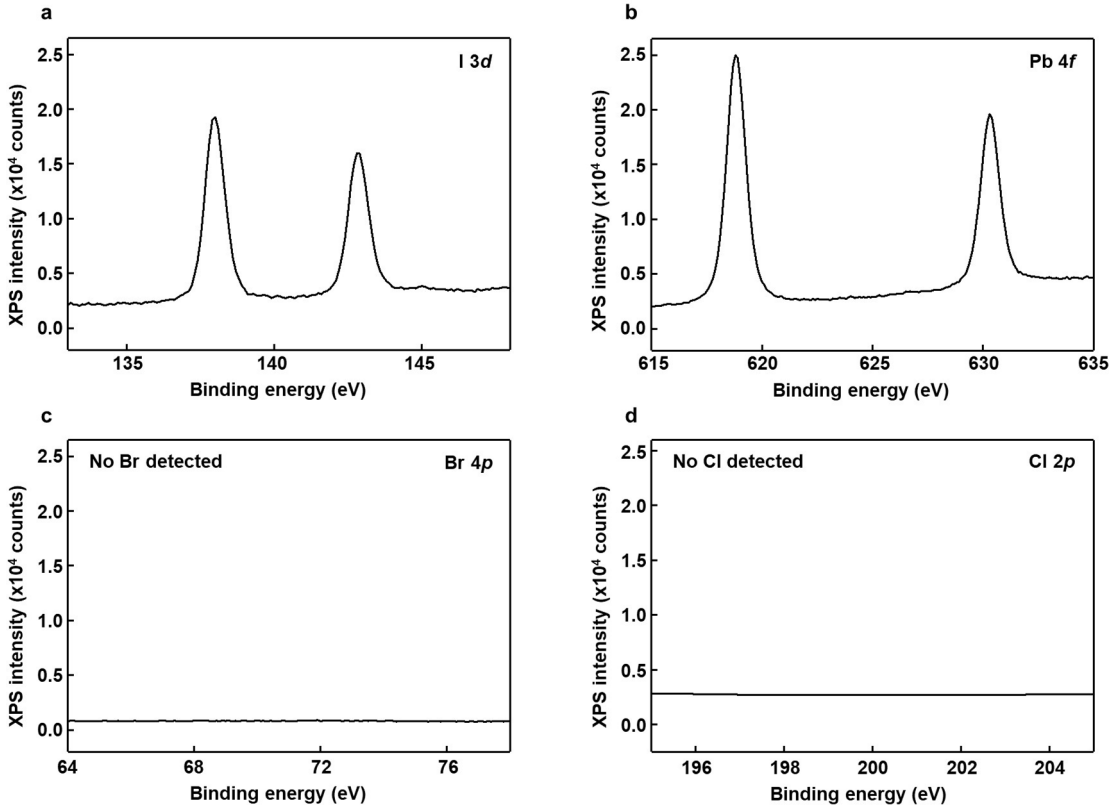
Extended Data Fig. 5 | Plastic strain relaxation study of the epitaxial α-FAPbI$_3$ thin films. a, **b**, Thickness-dependent in-plane XRD of −1.2% strained (**a**) and −2.4% strained (**b**) α-FAPbI$_3$ thin films. Vertical lines label the peak position of the fully strained films. Plastic strain relaxation at relatively high thickness can be evident by the peak shifting to lower angle and peak broadening. **c**, Thickness-dependent relaxation constant *R* of the epitaxial α-FAPbI$_3$ thin films with different strains. Results show that the critical thickness decreases with increasing strain. **d**, Fitting of the experimental critical thicknesses with the People and Bean and the Matthew and Blakeslee models (see Supplementary Information refs 69 and 70). Experimental results agree well with the People and Bean model, indicating that the plastic strain relaxation due to the dislocations generated during the epitaxial growth is the dominating relaxation mechanism.



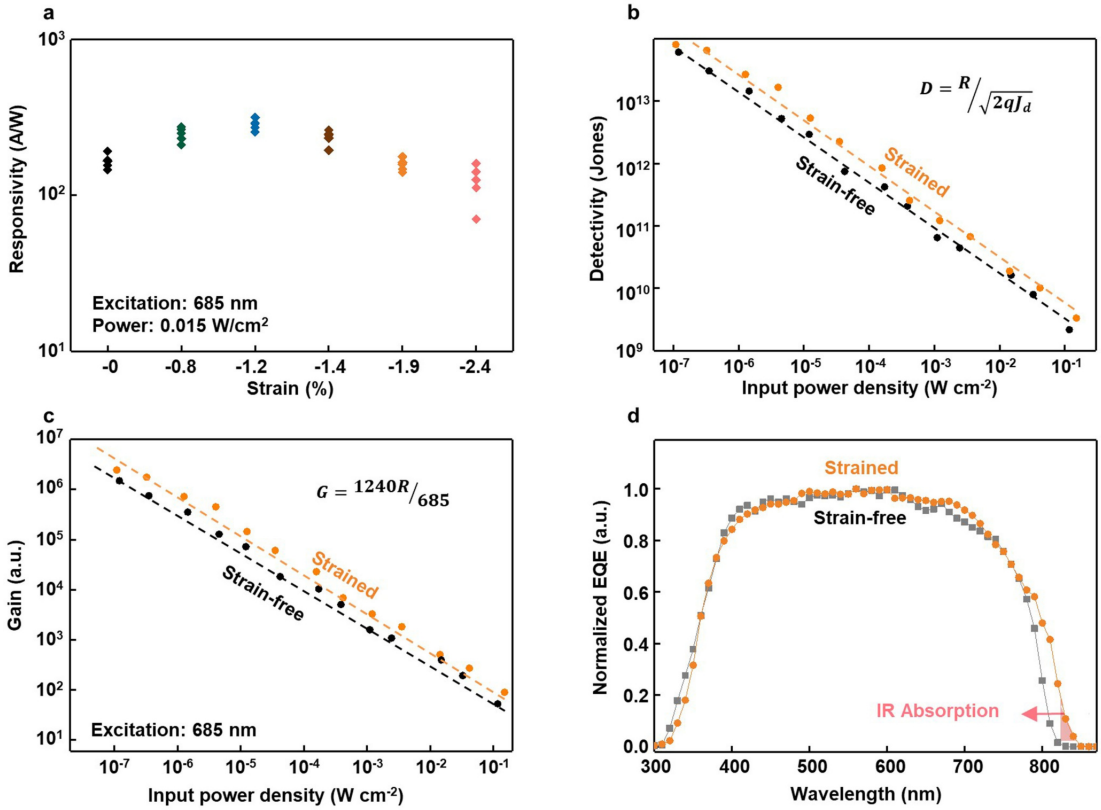
Extended Data Fig. 6 | UPS spectra of α-FAPbI$_3$ under different strains. a, The intersects of the curves with the baseline in the high-binding-energy region give the Fermi level position of corresponding strained α-FAPbI$_3$ films. There is a clear shift of the intersects to higher-binding-energy levels when the compressive strain becomes larger. **b**, The intersects of the curves with the baseline in the low-binding-energy region give the energy difference between the Fermi level and the VBM. All α-FAPbI$_3$ films have p-type character according to the calculated Fermi level position in the bandgap. Meanwhile, the VBM is pushed up more than the CBM with increasing strain. Inset, a schematic band diagram of the α-FAPbI$_3$ under different strains.



Extended Data Fig. 7 | Space-charge-limited-current measurement of the epitaxial α-FAPbI$_3$ with different strains. a–**d**, *I*–*V* characteristic curves for the space-charge-limited-current measurement of the epitaxial α-FAPbI$_3$ film with different strains. While the forward scans indicate a typical trap-filling process with increasing the applied voltage, the reverse scan doesn't show a detrapping process. Number of experiments, $n = 5$ for each strain value. n_{trap}, calculated trap density.



Extended Data Fig. 8 | XPS spectra of strained α-$FAPbI_3$. XPS spectra of: **a**, I 3*d*; **b**, Pb 4*f*; **c**, Br 4*p*; and **d**, Cl 2*p* photoelectrons from a strained α-$FAPbI_3$ film. Results show that Br and Cl are absent in the strained α-$FAPbI_3$.



Extended Data Fig. 9 | Photoconductor-type photodetector characterizations with a 685-nm laser. a, Responsivity as a function of strain level in α-$FAPbI_3$. Devices under −0.8%, −1.2% and −1.4% compressive strain give better responsivity compared to the strain-free devices. Further increasing the compressive strain can lead to a higher density of dislocations, which reduces the responsivity. Number of experiments, $n = 5$ for each strain value. **b**, **c**, Detectivity (**b**) and gain G (**c**) of the photodetector based on α-$FAPbI_3$ under −1.2% strain, indicating enhanced performance. q, element charge. J_d, dark current density. **d**, Normalized external quantum efficiency (EQE) of the photodetector based on α-$FAPbI_3$ under −1.2% strain, showing an extended infrared absorption range.

Article

Single-chain heteropolymers transport protons selectively and rapidly

https://doi.org/10.1038/s41586-019-1881-0

Received: 31 December 2018

Accepted: 1 October 2019

Published online: 8 January 2020

Tao Jiang[1,11], Aaron Hall[1], Marco Eres[1,2], Zahra Hemmatian[3,12], Baofu Qiao[4], Yun Zhou[5], Zhiyuan Ruan[1], Andrew D. Couse[1,13], William T. Heller[6], Haiyan Huang[7], Monica Olvera de la Cruz[4,8], Marco Rolandi[3] & Ting Xu[1,2,9,10]*

Precise protein sequencing and folding are believed to generate the structure and chemical diversity of natural channels[1,2], both of which are essential to synthetically achieve proton transport performance comparable to that seen in natural systems. Geometrically defined channels have been fabricated using peptides, DNAs, carbon nanotubes, sequence-defined polymers and organic frameworks[3–13]. However, none of these channels rivals the performance observed in their natural counterparts. Here we show that without forming an atomically structured channel, four-monomer-based random heteropolymers (RHPs)[14] can mimic membrane proteins and exhibit selective proton transport across lipid bilayers at a rate similar to those of natural proton channels. Statistical control over the monomer distribution in an RHP leads to segmental heterogeneity in hydrophobicity, which facilitates the insertion of single RHPs into the lipid bilayers. It also results in bilayer-spanning segments containing polar monomers that promote the formation of hydrogen-bonded chains[15,16] for proton transport. Our study demonstrates the importance of the adaptability that is enabled by statistical similarity among RHP chains and of the modularity provided by the chemical diversity of monomers, to achieve uniform behaviour in heterogeneous systems. Our results also validate statistical randomness as an unexplored approach to realize protein-like behaviour at the single-polymer-chain level in a predictable manner.

The monomer selection was based on previously designed RHPs[14]. The two hydrophobic monomers, methyl methacrylate (MMA) and 2-ethylhexyl methacrylate (EHMA), promote RHP insertion into lipid bilayers. Oligo(ethylene glycol) methyl ether methacrylate (OEGMA; number average molecular weight, $M_n \approx 500$ Da) was included to tailor the overall hydrophilicity and promote hydrogen-bonded chain (HBC) formation[15,16] within the lipid bilayer for proton transport. 3-sulfopropyl methacrylate potassium salt (SPMA) was used to reduce the aggregation propensity of the RHPs.

Although statistically random on a whole-chain level, each RHP chain was composed of segments with a range of cumulative hydrophobicities. The RHP sequences were generated using Compositional Drift, a program based on the Mayo–Lewis model, which considers the relative concentrations and the reactivity ratios of each comonomer pair, and on global monomer conversion[17]. Figure 1a shows 20 representative sequences of the RHP (termed 'RHP1') with a feeding compositional ratio of MMA:OEGMA:EHMA:SPMA = 5:2.5:2:0.5. Although the RHP1 sequences are statistically random, the heterogeneity is obvious in the cumulative hydrophobicity among different local segments within and among RHP1 chains. The local hydrophobicity along an RHP1 chain can be evaluated on the basis of the average hydrophilic–lipophilic balance (HLB) value of a sliding window (Fig. 1b)[18]. An RHP chain can be broken down into three types of segments according to the cumulative HLB value: those that prefer water, the lipid bilayer and the interface (Fig. 1c). With the same ensemble composition and statistical monomer distribution, individual RHP chains feature different sequences. However, those types of segment are seen for all calculated RHP1 sequences. The insertion of RHP1 into lipid bilayers was confirmed by fluorescence imaging and differential scanning calorimetry (DSC) measurements. Figure 1d shows the confocal images of liposomes incubated with RHP1 end-labelled with Texas Red dye. After washing off the excess polymer, the fluorescence signal from RHP1 was detected near the lipid membranes and overlapped with the signal from a premixed membrane-specific dye (NBD-PE). The DSC results show the broadening of the lipid phase transition in the RHP1-containing liposomes, confirming RHP1 insertion, rather than only surface absorption (Extended Data Fig. 1a, b).

We ran all-atom explicit solvent molecular dynamics simulations to visualize the distributions of RHP1 within a lipid bilayer (Fig. 1e,

[1]Department of Materials Science and Engineering, University of California Berkeley, Berkeley, CA, USA. [2]Department of Chemistry, University of California Berkeley, Berkeley, CA, USA. [3]Department of Electrical and Computer Engineering, University of California Santa Cruz, Santa Cruz, CA, USA. [4]Department of Materials Science and Engineering, Northwestern University, Evanston, IL, USA. [5]Division of Biostatistics, University of California Berkeley, Berkeley, CA, USA. [6]Neutron Scattering Division, Oak Ridge National Laboratory, Oak Ridge, TN, USA. [7]Department of Statistics, University of California Berkeley, Berkeley, CA, USA. [8]Department of Chemistry, Northwestern University, Evanston, IL, USA. [9]Materials Science Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [10]Kavli Energy NanoScience Institute, University of California Berkeley, Berkeley, CA, USA. [11]Present address: Department of Chemistry, The MOE Key Laboratory of Spectrochemical Analysis and Instrumentation, Xiamen University, Xiamen, China. [12]Present address: Quantapore Inc., San Francisco, CA, USA. [13]Present address: Department of Chemistry, Indiana University Bloomington, Bloomington, IN, USA. *e-mail: tingxu@berkeley.edu

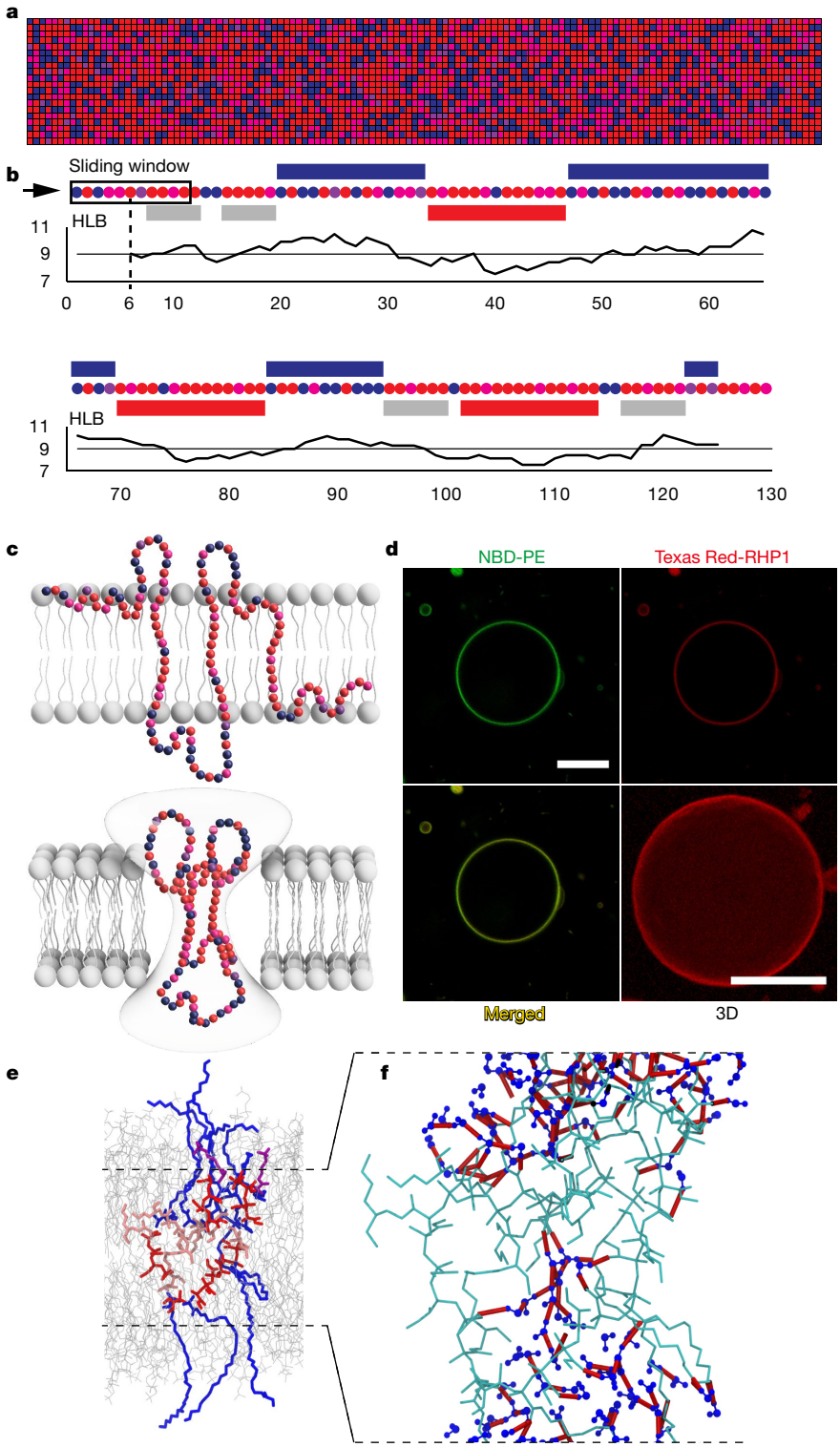


Fig. 1 | Segmental heterogeneity in RHP. a, Simulated sequences of 20 RHP1 chains (degree of polymerization, DP = 130). MMA, EHMA, OEGMA and SPMA are shown in red, pink, blue and purple, respectively. **b**, Hydrophilic (blue), long (red) and short (grey) hydrophobic segments in an RHP1 chain. Average HLB values for a sliding window of 11 monomers versus the monomer position are shown by the black line. **c**, Two-dimensional (2D) and three-dimensional (3D) illustrations of an RHP1 in a lipid bilayer. **d**, Confocal fluorescence images and 3D image reconstruction of an NBD-PE-labelled POPC liposome after incubation with Texas Red-labelled RHP1. Scale bar, 20 μm. **e**, Snapshot of the spatial distribution of an RHP1 chain (the first sequence in Extended Data Fig. 2a) in the lipid bilayer in the all-atom molecular dynamics simulation. **f**, Snapshot of the HBCs (red sticks) formed within a distance of 3.5 Å of the RHP1.

Extended Data Fig. 2). Most hydrophilic side chains are found to protrude into the water or near the membrane surfaces. Hydrophobic segments are distributed within the bilayer, anchoring a few OEG side chains to the nonpolar region of the bilayer. Those segments are rather static, and the OEG side chains in the bilayer move at slower rates than the ones in water (Supplementary Videos 1 and 2). The inserted RHP1 promotes the formation of rather dynamic HBCs among OEGs, methacrylate carbonyls and waters in the bilayer without adopting uniform three-dimensional structures (Fig. 1f, Supplementary Videos 3 and 4) and should act as 'stepping stones' to assist proton transport across the bilayers.

A bioprotonic device (Fig. 2a) was used to probe proton transport through a supported lipid bilayer (SLB) composed of POPE and POPG (see Methods; POPE:POPG molar ratio, 3:1) on top of a Pd/PdH_x contact, which translates a proton current into an electronic current[19]. With −200 mV (versus Ag/AgCl) applied to the Pd contact, incorporating RHP1 (RHP1:lipid molar ratio, 0.0005) into the SLB increased the current of H^+ across the SLB with respect to the SLB alone. After H^+ permeated

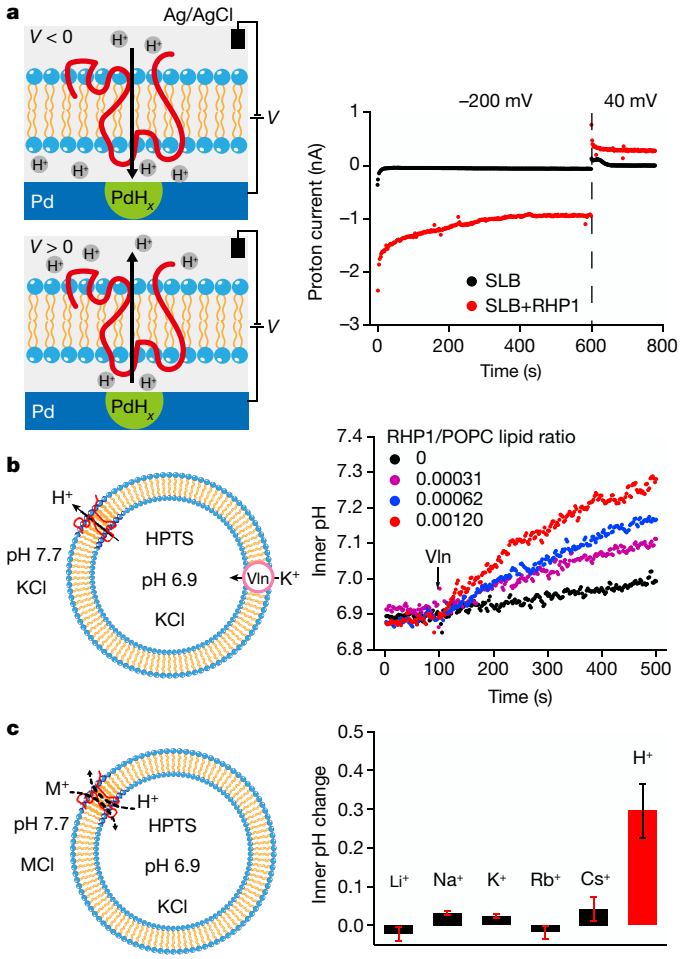


Fig. 2 | Transmembrane proton transport. a, Voltage-driven proton transport through RHP1 incorporated in an SLB on the Pd/PdH_x device. **b**, pH gradient-driven proton flux through RHP1 in the liposome. The arrow indicates the time when Vln was added. **c**, Inner pH changes to external alkali metal cations after an incubation time of 200 s without Vln (black bars). The red bar denotes the pH change with Vln and external K^+. Error bars are 1 s.d. ($n = 3$). The RHP1-to-lipid molar ratio is 0.0017.

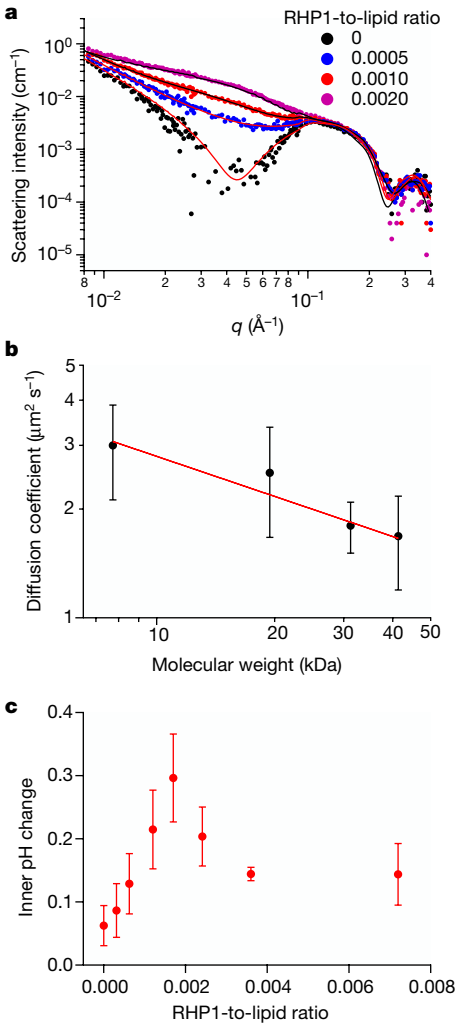


Fig. 3 | Interaction of RHP1 and the lipid bilayers. a, SANS profiles of d54-DMPC liposomes containing RHP1(~19 kDa) in 86% D_2O:H_2O. The solvent signal was subtracted. Solid lines denote a fit using a three-layer vesicle model for the liposome and a sphere model for the RHP1. **b**, Diffusion coefficients of the Texas Red-labelled RHP1s in the liposome membrane as a function of polymer molecular weight (7, 19, 30 and 40 kDa). Error bars are 1 s.d. ($n = 6$). **c**, Inner pH changes of the liposomes incubated with RHP1 over 200 s after adding Vln. Error bars are 1 s.d. ($n = 3$).

through the SLB, it was reduced at the Pd surface into H, which physisorbed onto the Pd surface to form PdH_x. When the voltage on the Pd contact was switched from −200 mV to +40 mV, the positive voltage oxidized H into H^+ and resulted in a positive proton current, i_{H+}, across the SLB. A higher positive i_{H+} was also detected for the RHP1 containing SLB. Using these data, we calculated that adding RHP1 increased the apparent membrane permeability of the SLB from 8.1×10^{-5} s^{-1} to 1.6×10^{-2} s^{-1}. The results confirm that RHP1 effectively promotes proton transport across the SLB.

The proton transport was further evaluated using a liposome-based fluorescence assay, in which the proton permeation was monitored using a pH-sensitive fluorophore pyranine (HPTS). The pH change was recorded from liposomes composed of the lipid POPC (see Methods; Fig. 2b) or the lipids POPE and POPG (POPE:POPG molar ratio, 3:1) (Extended Data Fig. 3a, b). Driven by a pH gradient of about 0.8, enhancement in the proton flux was observed by adding RHP1 to the liposome solutions, which was triggered by the K^+ ionophore valinomycin (Vln) that dissipates membrane potentials. To quantify the proton flux rate per RHP1 chain, we premixed the lipid and RHP1 before liposome preparation and went through freeze–thaw cycles to promote RHP1 insertion. We assumed that all RHP1 chains contribute to the proton transport and the calculated transport value represents the lower limit of the RHP1 performance. The stopped-flow fluorescence measurements show an initial proton flux per RHP1 of 8.1 ± 1.0 H^+ s^{-1} under these conditions, corresponding to a single-RHP1 conductance value of $(4.4 \pm 0.57) \times 10^{-17}$ S ($n = 3$, s.d.) (Extended Data Fig. 4a–d). This is comparable to the value reported for the proton-selective channel influenza virus M2, 2.9×10^{-17} S to 4.4×10^{-17} S, at pH 5.7 using a similar assay[20]. As a control, the conductance of the proton/cation channel gramicidin A (GramA) was measured to be $(4.0 \pm 0.43) \times 10^{-16}$ S ($n = 3$, s.d.), in the same magnitude as the reported values of 1.2×10^{-16} S to 2.1×10^{-16} S (refs. [21,22]). Slowed proton transport was observed after substituting D_2O for H_2O in this assay (Extended Data Fig. 4e). The isotope effect value of RHP1 (1.9) is larger than that of water pore channels (1.6), where protons are transported along water wires via the Grotthuss mechanism[13]. The large isotope effect agrees with the proposed transport pathway involving both RHP1 and water, rather than RHP1 forming a pore in the bilayer[13,23].

The insertion of RHP1 did not increase the water permeability of the bilayer, confirming the absence of distinct membrane leakage (Extended Data Fig. 4f). The direction of proton transport was varied to eliminate the snorkelling effect from the OEG side chains

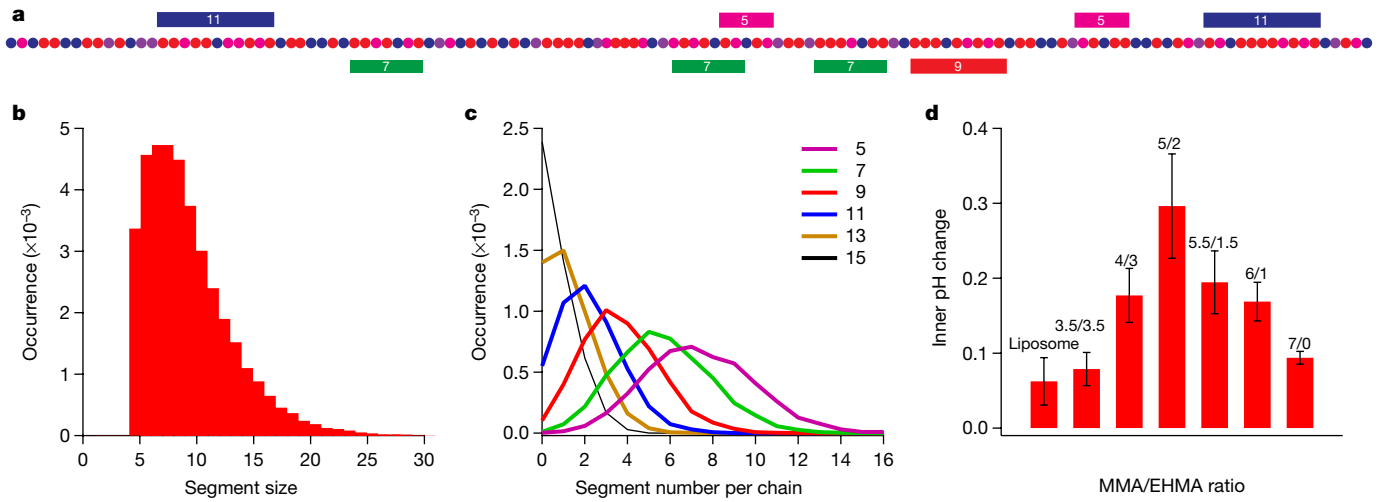


Fig. 4 | Sequence analyses of the RHPs. a, One set of hydrophobic segments containing one OEGMA in an RHP1 chain (DP = 130). **b**, Statistical size distribution of the hydrophobic segments containing one OEGMA in 4,500 simulated RHP1 sequences. **c**, Number of such segments per polymer chain, with the minimum segment size ranging from 5 to 15. Segments of even numbers are omitted for clarity. **d**, Transport performance of the RHP variants, evaluated using the liposome-based fluorescence assay (Fig. 2b). Inner pH changes of the RHP-containing liposomes (RHP:lipid ratio of 0.0017) over 200 s after adding Vln. Error bars are 1 s.d. (n = 3).

of bilayer-surface-absorbed RHP1 chains (Extended Data Fig. 3c). In addition, an RHP1 derivative (RHP1,100) containing an OEGMA with a longer side chain ($M_n \approx 1{,}100$ Da) was tested. Although it favours bilayer-surface binding, RHP1,100 did not promote proton permeation under the same conditions (Extended Data Figs. 1c, d, 3b).

Small-angle neutron scattering (SANS) was used to characterize the RHP1 size and oligomeric state in a lipid bilayer (Fig. 3a, Extended Data Fig. 5). Owing to the large scattering length density (SLD) contrast against RHP1, d54-DMPC (see Methods) was used to differentiate the signal of RHP1 from that of the lipid background while still providing RHP1 proton transport activity (Extended Data Fig. 3d). The SANS measurements provided an average RHP1 (about 19 kDa) radius of 2.2 nm. The physical volume occupied by one dry RHP1 chain was calculated to be about 32 nm^3, assuming a polymer density of 1.0 g ml^{-1}. Using a sphere model for single RHP1 chains, the SANS results lead to an average particle size of about 45 nm^3. Thus, the RHP1 chains formed a unimolecular assembly under the experimental conditions used. Fluorescence recovery after photobleaching (FRAP) measurements were performed on the liposome surface using RHP1s with molecular weights ranging from ~7 kDa to ~40 kDa (Fig. 3b). Despite large fluctuations in the measured diffusion rate (D), the correlation between D and the RHP1 molecular weights further excludes extensive polymer aggregation in the lipid bilayer. Additionally, in the proton transport study, promoted proton transport was observed with RHP1-to-lipid ratios increasing from 0 to 0.0017 (RHP1 concentration, 0–48 nM) (Fig. 3c). The trend of the pH increase is reversed at higher RHP1 concentration, probably owing to the formation of RHP1 oligomers that were excluded from the lipid bilayer. Together, these results confirmed that the monomeric form of RHP1 is the active species for the polymer concentration window reported here.

Diffusion of alkali metal cations across a lipid bilayer requires a pore size larger than the size of the dehydrated cations, whereas protons can either diffuse through pores in the form of water–proton complexes or be transported without pores by hopping along HBCs[15,16]. The ionic selectivity of RHP1-based transport against other alkali metal cations (including Li^+, Na^+, K^+, Rb^+ and Cs^+) was studied using an external cation exchange approach[10] (Fig. 2c). In the presence of both proton and cation gradients across the bilayer, no distinct co-transport of protons and alkali metal ions was observed, which was manifested by minimal pH changes in the interior of the liposomes. Such insensitivity indicates that RHP1-based transport is highly selective towards protons. This is consistent with the proposed HBC-based transport mechanism in the transmembrane RHP1 segments and the unimolecular nature of RHP1 within the bilayer.

An important hallmark of RHPs is their sequence heterogeneity. Even with the same ensemble composition and statistical monomer distribution, individual RHP chains feature different sequences. However, the statistical population and distribution of segments according to cumulative hydrophobicity should remain the same. Despite the heterogeneity, RHP1 shows consistent transport performance. This may be attributed to the enhanced adaptability that is unavailable in more defined systems, and may be advantageous for achieving efficient molecular transport. Analyses were performed to gain insight into the interplay between differences in monomer sequences and similarities in the statistical structure of segments (that is, similar statistical parameters, which characterize the heterogeneity of segments) and how these may affect the RHP1 arrangement in lipid insertion and proton transport. Segments rich in MMA and EHMA can be overall hydrophobic while still including a few OEGMA monomers, and sufficiently long to span across lipid bilayers. Similarly, there exist segments rich in hydrophilic monomers that include a few hydrophobic units, which generally prefer the aqueous phase. Both are chemically heterogeneous but with distinctive cumulative hydrophobicities. Within each RHP1 chain, there are also abundant segments that cannot be easily assigned as overall hydrophobic or hydrophilic. Rather, their cumulative chemical features vary substantially depending on the choice of segment length and ending position. Nevertheless, they appear to be generally amphiphilic and will probably have a short residence time in either phase. Their rapid kinetics of insertion and desorption from bilayers, instead of a static state, may also facilitate proton transport.

We hypothesize that the inclusion of a small amount of OEGMA in the lipid bilayer core is a key requirement for the observed proton transport. Analysing the hydrophobic segments including OEGMA can be done with various criteria over the block length, number of OEGMAs included and flanking monomer sequence. As one example, Fig. 4a highlights the hydrophobic segments containing one OEGMA in a representative RHP1 chain. An in-house-developed program was used to evaluate the length distribution of such segments in a total of 4,500 RHP1 (degree of polymerization, DP = 130) sequences generated by Compositional Drift, considering compositional dispersity

(Extended Data Fig. 6a). As seen in Fig. 4b, RHP1 contains the segments that fulfil the length requirement to span the lipid bilayers. The core of a typical lipid bilayer (~2.7 nm for a POPC bilayer[24]) can theoretically accommodate 11 MMA-based units in their fully extended form. Within a certain length range, shorter and longer segments should also be capable of integrating into the bilayer. RHP1 segments and surrounding lipids can adjust their chain conformation and/or bilayer thickness, respectively, to adapt to hydrophobic mismatch. Further analysis, shown in Fig. 4c, suggests that there are a limited number of such segments per RHP1 chain (DP = 130). For example, an average of three segments per polymer chain was counted with a segmental length of ≥9, and two segments per chain with length ≥11. RHP1 length dispersity was also considered in the segment analyses. The statistical distribution of the segment size is less affected by the polymer length variation once the RHP1 is sufficiently long (DP ≥ 40) (Extended Data Fig. 6b, c), although the segment numbers per chain have a positive correlation with the RHP1 chain length (Extended Data Fig. 6d). We also used the probabilistic framework of the hidden Markov model (HMM) to predict the transmembrane segments of RHP1 (Extended Data Fig. 7). The HMM investigation agrees with the aforementioned sequence analysis, indicating that RHP1 chains are rich in transmembrane segments and include those that allow OEGMAs to embed in the bilayer core to facilitate proton transport. Such hydrophobic segments are limited in their population for the molecular weight range studied here, thus forming a narrow crossing path for proton transport.

According to the information gained, segmental heterogeneity afforded by four distinct monomers appears to be a key design feature that is necessary to capture several critical features of natural channels with regards to proton transport. Experimentally, none of the three-monomer-based RHPs shows efficient transport activity (Extended Data Fig. 3b). It is reasonable to believe that a statistical design approach should also lead to other RHPs with proton transport properties. We synthesized a series of RHP variants, varying the MMA:EHMA ratio from 3.5:3.5 to 7:0. The overall fraction of these two hydrophobic monomers remains the same in order to maintain similar segmental distributions. The RHPs within the compositional range of 4:3 to 6:1 MMA:EHMA could promote proton transport, albeit with different performances (Fig. 4d). RHP1 (MMA:EHMA = 5:2) elicits the best proton permeation, whereas either increasing or decreasing the portion of EHMA reduces the overall transport performance. Shifting the MMA:EHMA ratio should alter the capability of polymer insertion into lipid bilayers by changing the overall polymer hydrophobicity/hydrophilicity and the balance between polymer–polymer and polymer–lipid interactions. Decreasing the MMA:EHMA ratio increases the occurrence of EHMA monomers distributed adjacent to each other. With a bulky ethyl hexyl side-chain, local abundance of EHMA would also sterically hinder the interplay between the polymer carbonyl and water/OEGMA, affecting the proton transport efficiency (Extended Data Fig. 6e, 6f).

The present study confirms that exploring and manipulating the segmental heterogeneity in RHPs at the single-chain level is feasible and effective to mimic protein-like behaviour. Considering the vast number of monomers available and recent advances in polymer chemistry, the possibility to generate RHP-based functional materials[14,25–27] and eventually marry the synthetic and biological fields are almost unlimited.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1881-0.

1. Stouffer, A. L. et al. Structural basis for the function and inhibition of an influenza virus proton channel. *Nature* **451**, 596–599 (2008); corrigendum **452**, 380 (2008).
2. Schnell, J. R. & Chou, J. J. Structure and mechanism of the M2 proton channel of influenza A virus. *Nature* **451**, 591–595 (2008).
3. Ghadiri, M. R., Granja, J. R. & Buehler, L. K. Artificial transmembrane ion channels from self-assembling peptide nanotubes. *Nature* **369**, 301–304 (1994).
4. Xu, T. et al. Subnanometer porous thinfilms by the co-assembly of nanotube subunits and block copolymers. *ACS Nano* **5**, 1376–1384 (2011).
5. Joh, N. H. et al. De novo design of a transmembrane Zn^{2+}-transporting four-helix bundle. *Science* **346**, 1520–1524 (2014).
6. Langecker, M. et al. Synthetic lipid membrane channels formed by designed DNA nanostructures. *Science* **338**, 932–936 (2012).
7. Hinds, B. J. et al. Aligned multiwalled carbon nanotube membranes. *Science* **303**, 62–65 (2004).
8. Geng, J. et al. Stochastic transport through carbon nanotubes in lipid bilayers and live cell membranes. *Nature* **514**, 612–615 (2014).
9. Percec, V. et al. Self-assembly of amphiphilic dendritic dipeptides into helical pores. *Nature* **430**, 764–768 (2004).
10. Weiss, L. A., Sakai, N., Ghebremariam, B., Ni, C. Y. & Matile, S. Rigid rod-shaped polyols: functional nonpeptide models for transmembrane proton channels. *J. Am. Chem. Soc.* **119**, 12142–12149 (1997).
11. Vial, F., Oukhaled, A. G., Auvray, L. & Tribet, C. Long-living channels of well defined radius opened in lipid bilayers by polydisperse, hydrophobically-modified polyacrylic acids. *Soft Matter* **3**, 75–78 (2007).
12. Jung, M., Kim, H., Baek, K. & Kim, K. Synthetic ion channel based on metal–organic polyhedra. *Angew. Chem.* **47**, 5755–5757 (2008).
13. Si, W. et al. Selective artificial transmembrane channels for protons by formation of water wires. *Angew. Chem.* **50**, 12564–12568 (2011).
14. Panganiban, B. et al. Random heteropolymers preserve protein function in foreign environments. *Science* **359**, 1239–1243 (2018).
15. Nagle, J. F. & Morowitz, H. J. Molecular mechanisms for proton transport in membranes. *Proc. Natl Acad. Sci. USA* **75**, 298–302 (1978).
16. Decoursey, T. E. Voltage-gated proton channels and other proton transfer pathways. *Physiol. Rev.* **83**, 475–579 (2003).
17. Smith, A. A. A., Hall, A., Wu, V. & Xu, T. Practical prediction of heteropolymer composition and drift. *ACS Macro Lett.* **8**, 36–40 (2019).
18. Kyte, J. & Doolittle, R. F. A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132 (1982).
19. Hemmatian, Z. et al. Electronic control of H^+ current in a bioprotonic device with Gramicidin A and Alamethicin. *Nat Commun.* **7**, 12981 (2016)
20. Lin, T. I. & Schroeder, C. Definitive assignment of proton selectivity and attoampere unitary current to the M2 ion channel protein of influenza A virus. *J. Virol.* **75**, 3647–3656 (2001).
21. Krishnamoorthy, G. Temperature jump as a new technique to study the kinetics of fast transport of protons across membranes. *Biochemistry* **25**, 6666–6671 (1986).
22. Tunuguntla, R. H., Allen, F. I., Kim, K., Belliveau, A. & Noy, A. Ultrafast proton transport in sub-1-nm diameter carbon nanotube porins. *Nat. Nanotechnol.* **11**, 639–644 (2016).
23. DeCoursey, T. E. & Cherny, V. V. Deuterium isotope effects on permeation and gating of proton channels in rat alveolar epithelium. *J. Gen. Physiol.* **109**, 415–434 (1997).
24. Kučerka, N., Tristram-Nagle, S. & Nagle, J. F. Structure of fully hydrated fluid phase lipid bilayers with monounsaturated chains. *J. Membr. Biol.* **208**, 193–202 (2006).
25. Golumbfskie, A. J., Pande, V. S. & Chakraborty, A. K. Simulation of biomimetic recognition between polymers and surfaces. *Proc. Natl Acad. Sci. USA* **96**, 11707–11712 (1999).
26. Mowery, B. P. et al. Mimicry of antimicrobial host-defense peptides by random copolymers. *J. Am. Chem. Soc.* **129**, 15474–15476 (2007).
27. Tribet, C., Audebert, R. & Popot, J. L. Amphipols: polymers that keep membrane proteins soluble in aqueous solutions. *Proc. Natl Acad. Sci. USA* **93**, 15047–15050 (1996).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Materials

Chemicals were purchased from Sigma Aldrich and Fisher Scientific unless otherwise noted. Phospholipids were purchased from Avanti Polar Lipids Inc., including 1-palmitoyl-2-oleoyl-sn-glycero-3-phosphoethanolamine (POPE), 1-palmitoyl-2-oleoyl-sn-glycero-3-phospho-(1′-rac-glycerol) (sodium salt) (POPG), 1-palmitoyl-2-oleoyl-glycero-3-phosphocholine (POPC), 1,2-dimyristoyl-sn-glycero-3-phosphocholine (DMPC) and 1,2-dimyristoyl-d54-sn-glycero-3-phosphocholine (d54-DMPC). Water was Milli-Q-grade water (18.2 MΩ cm).

RHP syntheses

RHPs were synthesized as described previously[14]. Azobisisobutyronitrile was recrystallized from ethanol. Inhibitors were removed using cryodistillation (methyl methacrylate, 2-ethylhexyl methacrylate) or by passing over a column of neutral alumina (ethylene glycol methyl ether methacrylate ($M_n \approx 500$ g mol^{-1}). 3-sulfopropyl methacrylate potassium salt (98%), ethyl-2(phenylcarbanothioylthio)-2-phenylacetate (98%), trioxane and solvents were used as received. Polymerization solutions were prepared and transferred to 20-ml glass ampules. These solutions were subjected to four freeze–pump–thaw cycles and then flame-sealed at ~50 mtorr. Polymerizations were carried out at 80 °C for typically 6–12 h. The ampules were then quenched in liquid nitrogen and opened. Polymers were precipitated by dropwise addition to stirring pentane and then transferred to 3,500 molecular-weight cut-off centrifuge filters (Amicon), washed five times using water and lyophilized.

^{1}H nuclear magnetic resonance (NMR) spectra were obtained with either a Bruker Avance 400 spectrometer (400 MHz) using a 5-mm *Z*-gradient broadband observe probe, a Bruker Avance AV 500 spectrometer (500 MHz) using a *Z*-gradient triple broadband inverse detection probe, or a Bruker Avance AV 600 spectrometer (600 MHz) using a *Z*-gradient triple broadband inverse detection probe. Monomer conversion was measured on crude reaction mixtures in DMSO-d_6, using trioxane as an internal standard. The RHPs were characterized using gel permeation chromatography with an Agilent 1260 Infinity series instrument equipped with two Agilent PolyPore columns (300 mm × 7.5 mm), calibrated using poly(ethylene glycol) standards. Dimethylformamide with 0.05 M LiBr was used as eluent at 0.7 ml min^{-1} at 50 °C. Analyte samples at 2 mg ml^{-1} were filtered through 0.2-μm polytetrafluoroethylene membranes (VWR) before injection (20 μl). The physiochemical properties of the synthesized RHPs are listed in Extended Data Table 1.

Preparation of large unilamellar vesicles

2.5 mg of POPC or a mixture of POPE and POPG (3:1 molar ratio) was dissolved in chloroform and subjected to solvent removal to make a lipid film, which was dried in vacuum overnight. The lipids were rehydrated by a 0.5 ml inner buffer (10 mM HEPES-KOH, pH 6.9, 100 mM KCl) at 37 °C for 1 h. 0.5 mM pyranine was included in the inner buffer in the pyranine assay. The resulting suspension was subject to six cycles of freeze–thawing using a dry ice–ethanol bath and a 37 °C water bath. Large unilamellar vesicles (LUVs) were generated by extruding 11 times through a 400-nm polycarbonate membrane followed by a 200-nm polycarbonate membrane. Free dye was removed using a Sephadex G-25 desalting column (GE healthcare). LUV formation was confirmed using dynamic light scattering and transmission electron microscopy. The LUVs composed of POPE+POPG and of POPC only had diameters of 182 ± 16 nm and 189 ± 9 nm, respectively. The lipids were quantified using the Stewart method[28].

Preparation of giant unilamellar vesicles

Giant unilamellar vesicles (GUVs) for microscopy and FRAP investigations were prepared using a reported agarose-based hydration method[29]. 300 μl of 1% melted agarose solution (low melt; gelling temperature, $T_g \approx 26$ °C; melting temperature, $T_m \leq 65$ °C; electroendosmosis, EEO ≤ 0.12) was added to a cover glass (22 mm × 22 mm) and dried on a hotplate. 80 μg POPC in chloroform was evenly spread on the agarose film. The chloroform was removed under vacuum. The film was covered by a 1-ml HEPES buffer (10 mM HEPES-KOH, pH 6.9) to induce GUV growth for 0.5 h. Liposomes were released by gentle pipetting and immobilized into a thin layer of 0.5% agarose on a glass slide for microscopy investigations.

Interactions between RHPs and liposomes

Confocal fluorescence imaging. RHPs were end-modified with a Texas Red fluorescent dye via thiol-maleimide chemistry. GUVs doped with NBD-PE (100:1 molar ratio) were incubated with the Texas Red-labelled RHP1s (2,000:1 molar ratio) for 0.5 h. Unbound polymers were rinsed off right before imaging. The imaging was performed on a Zeiss LSM 880 with a 63× (numerical aperture, NA, 1.4) oil or 20× (NA, 1.0) water-immersion objectives and one Airy unit. Lasers at 488 nm and 594 nm were employed to excite the NBD and Texas Red, respectively. Fluorescence signals were collected in the 500–550 nm and 605–655 nm bands.

DSC investigation. DSC thermograms probing the lipid phase transition were acquired using a VP-MicroCal calorimeter (GE). The samples and buffers were equilibrated at 5 °C for 30 min and heated from 5 °C to 40 °C at a rate of 1 °C min^{-1}. The heating and cooling process was repeated for several cycles. The buffer or the RHP-containing buffers were used as the background. The full-width at half-maximum of the peak was obtained with baseline correction using the Origin software (Microcal).

FRAP measurement. FRAP experiments were conducted using Texas Red-labelled RHP1 containing-GUVs. Five prebleach images were taken at an attenuated laser intensity (0.5%, 561 nm). Subsequently, photobleaching was performed through a circular region of interest of nominal radius 3 μm and 20 bleaching iterations using the 514-nm and 561-nm laser lines at maximum intensity. After bleaching, recovery was monitored from the subsequent 95 frames with the excitation switched back to the attenuated intensity. A background scan and bleaching under the same conditions were conducted for FRAP curve correction. The recovery curves were fitted and the diffusion coefficient was obtained as reported previously[30].

Molecular dynamics simulation

All-atom explicit solvent molecular dynamics simulation. The simulations were performed using GROMACS (5.0.7)[31] with the CHARMM 36m force field[32,33] and the recommended TIP3P water model, the structure of which was constrained using the SETTLE algorithm[34,35]. RHP1 sequences (DP = 40) were created by varying the random seed with fixed component ratios. The initially extended RHP1 chains were first equilibrated in vacuum for 0.1 ns until they collapsed. Each RHP1 was then embedded at the centre of a bilayer with 120 POPC lipids. 7,200 water molecules were then added with approximately 0.1 M KCl. The initial structure (Extended Data Fig. 2d) was created using Packmol[36]. This process mimics the experimental procedures used to promote RHP1 insertion by premixing RHP1 and lipid and the freeze–thawing process used to create transient pores on the membranes.

The potential energy was minimized, followed by equilibration using the NPT ensemble (constant particle number, pressure and temperature). To relax the lipid bilayer, the *Z*-dimensional coordinates of all of the POPC atoms were constrained for 0.1 ns using a force constant of 1,000 kJ mol^{-1} nm^{-2}, followed by an exclusive constraint of the POPC phosphorous atoms for 0.1 ns. The equilibration simulation was performed at 310.15 K for 300 ns. In the production simulation at

298.15 K, the Nose–Hoover algorithm was employed to separately couple the temperatures of the polymer–POPC complex and the others. Semiisotropic pressure coupling ($P_{XY} = P_Z = 1$ bar; time constant, 5.0 ps; compressibility 4.5×10^{-5} bar^{-1}) was applied via the Parrinello–Rahman algorithm[37]. The neighbour search was performed up to 1.2 nm. The short-range Coulomb interactions were calculated up to 1.2 nm using the particle mesh Ewald method[38,39] for the long-range electrostatic interactions. The Lennard–Jones 12–6 potential was switched off from 0.8 nm to 1.2 nm with the force switching method. These parameters have been recommended for the accurate reproduction of the original CHARMM simulations on lipid membrane systems[40]. A simulation integration time step of 2 fs was employed, with all the hydrogen-involved covalent bond lengths constrained. Each production simulation was performed for 500 ns, with the final configuration presented in Extended Data Fig. 2d. The convergence of the simulations was confirmed by the area per lipid and the polymer–lipid interactions (Extended Data Fig. 2e). The HBCs were visualized using the package VMD[41].

Nine parallel simulations were performed with nine different polymer chains (Extended Data Fig. 2a). Five chains were found to span the bilayer roughly symmetrically (Fig. 1e, Extended Data Fig. 2b). The other four chains were found near the membrane–water interfaces. We believe that multiple factors influence the insertion of RHP1, including the overall monomer distributions along the RHP1 chain, the chain length and the transient RHP1 conformations. The sequence analysis showed that the average number of hydrophobic segments per RHP1 chain increases with longer chain lengths. For relatively short RHP1 chains (DP = 40; green line in Extended Data Fig. 6d), it is not surprising that not all of the nine different chains span through the lipid bilayer in these simulations.

Proton transport on LUVs

Transport on LUVs. With excitation at both 460 nm and 415 nm, the ratio of pyranine emission (I) at 510 nm (I_{ex460}/I_{ex415}) was monitored to probe the inner pH change over time. The tested fluorescence ratios of pyranine in the lumen of the LUVs were linearly proportional to pH values in the range 6.4–7.4. Purified LUV solutions were diluted ten times into the 1-ml sample with an outside buffer (10 mM HEPES-KOH pH 7.7, 100 mM KCl). Proper amounts of the RHP solutions at 0.1 or 1.0 mg ml^{-1} were added into the LUV solutions with continuous stirring. After incubation for 5 min, I_{460}/I_{415} was monitored over 500 s. 3 µl of 10 µM Vln in ethanol was added at 100 s to trigger the proton flux. To evaluate the ion selectivity of RHP1, purified solutions of RHP1-containing LUV were diluted ten times in 1 ml of outside buffers containing one type of alkali metal cation (10 mM HEPES-KOH pH 7.7, 100 mM MCl, M = Li, Na, K, Rb, Cs). Without Vln, changes in the inner pH were recorded over 200 s. The fluorescent measurements were performed on a Perkin Elmer KS55 fluorescence spectrometer and a micro-cuvette under constant stirring.

Transport investigation on stopped-flow fluorimeter

Proton transport rate. RHP1 and the lipids were premixed in chloroform before LUV preparation. Purified RHP1-containing LUV solutions were diluted 33 times with the inner solution and mixed with 30 nM Vln. 15 µl of the RHP1–Vln–LUV mixture in the inner solution was rapidly exposed to 15 µl of the outside solution using the AutoSF 120 stopped-flow fluorimeter (KinTek) at a flow rate of 8 ml min^{-1}. With excitation at 460 nm, the emission of the encapsulated pyranine was monitored as a function of time using a 520 nm/20 nm fluorescence filter. To evaluate the isotope effect of the proton transport, a similar proton transport experiment was performed by substituting D_2O for H_2O, including in the LUV preparation and purification processes. A deuterated base was used to titrate the pD. The pD value was measured using a glass pH electrode and corrected by adding 0.4 to the pH readout. The isotope effect value was obtained from the ratio of the conductance in the H_2O assay to that in the D_2O-based assay. The initial pH change rate was fitted to obtain the conductance of the channels or RHP1 and the permeability of membranes, using equations reported previously[22].

Water transport. LUVs were prepared in a similar way as for the proton transport assay, using 20 mM HEPES-KOH, pH 7.5, 100 mM KCl containing 10 mM carboxyfluorescein. The LUVs were purified using the same buffer without the dye. Using stopped-flow fluorescence at 7 °C, the channel- or RHP1-containing LUVs in this inner solution were rapidly exposed to an outside solution (20 mM HEPES-KOH, pH 7.5, 100 mM KCl, 100 mM sucrose) and subjected to a fluorescence measurement with excitation at 490 nm and the 520 nm/20 nm filter. The relative volume was linearly related to the relative fluorescence at the present conditions. The fluorescence quenching curves for the shrinking liposomes were fitted with a single-exponential function. The osmotic water permeability was calculated as described previously[42,43].

RHP conformation in liposomes

SANS investigation. SANS measurements were conducted at the GP-SANS instrument of the High Flux Isotope Reactor at Oak Ridge National Laboratory[44] and at the NG7 instrument of the National Institute of Standards and Technology. The sample-to-detector distances were set to generate a range of momentum transfer, q, from 0.004 to 0.6 Å^{-1}, where $q = 4\pi\sin\theta/\lambda$, where 2θ denotes the scattering angle from the incident beam. The wavelength λ is 5 Å for NG7 and 4.75 Å for GP-SANS. Small unilamellar vesicles (SUVs) composed of d54-DMPC were prepared as described above, using 86 vol% D_2O in H_2O and a 100-nm membrane. The SLD of the 86% D_2O theoretically matches the average SLD of d54-DMPC. Although 100% contrast match cannot be achieved owing to the deuterated lipid tail and H-based lipid head, using 86% D_2O dampens the SANS signals from the liposome. The SANS measurements were performed at 37 °C in 2-mm-path-length quartz cells using solutions of the SUVs. The 2D scattering results were azimuthally averaged to produce one-dimensional SANS curves. The data were modelled using a three-shell vesicle model for the RHP1-free SUVs and a sphere model for the free RHP1. A sum of these two models was used to fit the SANS profiles of the SUV–RHP1 samples. All data fitting was performed using the SASView program (version 4.2.0, http://www.sasview.org/).

Sequence analyses

HLB. The HLB value was used to evaluate the solubility of monomer side-chains through group contribution theory. Using the equation $\mathrm{HLB} = 7 + \sum_i n_i \mathrm{HLB}_i$, where n_i is the number of the ith chemical group in the molecule with corresponding value HLB_i. The HLB value for each monomer side chain was estimated as: HLB(MMA) = 8.45, HLB(EHMA) = 5.12, HLB(OEGMA) = 11.4 and HLB(SPMA) = 18.5. A lower HLB value denotes higher hydrophobicity and a higher value means greater hydrophilicity. A Python program was created to continuously calculate the average segmental HLB values for a window sliding from the alpha to the omega ends of the simulated RHP1 chains. The window advanced by one monomer each time. We used a span containing odd numbers of monomers, and assigned the average HLB value of that span to its middle monomer. For instance, for the window of 11 monomers shown in Fig. 1b, the average HLB value of the span from monomer 1 to monomer 11 was calculated and plotted at monomer position 6. The next HLB value, corresponding to the span from monomer 2 to monomer 12, was plotted at monomer position 7.

Hydrophobic segments. A Python program was used to analyse the length and number of hydrophobic segments containing one OEGMA in the RHPs. The hydrophobic embedded OEGMA was set to be two or more monomers away from the ends of the segments. The number of uninterrupted hydrophobic monomers on both sides of an OEGMA was counted.

Segment prediction

HMM prediction. The HMM has a successful history in describing latent statistical structures and incorporating various constraints, despite its limitation in duration modelling. Given an observable output sequence of length T (for example, $y = (y_1, y_2, \ldots, y_T)$), the basic goal for an HMM[45] is to infer the corresponding hidden state sequence of the same length, $q = (q_1, q_2, \ldots, q_T)$. In our study, y_i denotes the monomers and q_i represents the states defined above: state 1, state 2 and state 3. The HMM parameters include:

1. The initial probabilities of each state, $\Delta = \{\pi_i\}$, where $\pi_i = P(q_1 = i)$ and i is the index for the three different hidden states (that is, $i = 1, 2, 3$).

2. The transition probability matrix $A = \{a_{ij}\}$, where $a_{ij} = P(q_{t+1} = j \,|\, q_t = i)$, that is, the probability of next step being in state j ($j = 1, 2, 3$) if currently it is in state i.

3. The emission probability matrix $B = \{b_{ij}\}$ is defined as $b_{ij} = P(y_t = j | q_t = i)$, where j is the index for the four different types of monomers (that is, $j = 1, 2, 3, 4$).

Then the joint probability of states (that is, $q = (q_1, q_2, \ldots, q_T)$) and observed outputs (that is, $y = (y_1, y_2, \ldots, y_T)$) can be written as:

$$P(y, q) = \pi_{q_1} b_{q_1 y_1} \left(\prod_{t=2}^{T-1} b_{q_t y_t} a_{q_t q_{t+1}} \right) b_{q_T y_T}$$

The Baum–Welch algorithm[46] is typically used to find the unknown parameters Δ, A and B, and the Viterbi algorithm[47] is used to determine the optimal state path for each observed sequence.

The unobserved states represent the three possible regions on RHP1s that prefer water, the lipid–water interface, or the core of lipid bilayers, and states 1, 2 and 3 denote the corresponding hidden states. We implemented a third-order HMM in which (i) the transitions between states were modelled by a third-order Markov chain and (ii) third-order emission probabilities were used to characterize the monomer preferences in different RHP1 regions and to model the dependencies between neighbouring monomers—that is, the probability of generating a monomer at a position would depend not only on the hidden state at that position, but also on the states at the previous two positions. Additionally, we incorporated the following rules in the HMM to better approximate reality and facilitate parameter inference (Extended Data Fig. 7b):

1. Despite the ambiguous criteria for defining a state-3 segment, we began with segments full of hydrophobic monomers and hydrophobic segments containing one OEGMA.

2. To differentiate between OEGMAs included in a state-3 segment and those adjacent to such segments, the OEGMA included was set to be at least two monomers away from the ends of the segments.

3. Owing to the high polarity, SPMA was fixed to appear only in state 1 in the current setup.

4. Hydrophobic segments (states 2 and 3) end with hydrophilic monomers (OEGMA and/or SPMA). Those boundary monomers, if not included in an extended state-3 segment (that is, the OEGMA within a long hydrophobic segment), were considered to be in state 1. We note that when a length constraint for state 3 has to be incorporated (as shown in Extended Data Fig. 7b), the design of the HMM would be more challenging and the HMM might not perfectly represent the observed data distribution.

Proton transport on SLBs

Transport on an SLB. SUVs were prepared in a similar way to the LUVs, except for using a 100-nm extrusion membrane. 150 µl of 2.5 mg ml^{-1} SUVs was deposited on a clean Pd contact and incubated overnight with mild agitation to facilitate formation of the SLB. Unfused SUVs were washed off using the liposome buffer. The SLB formation was verified by liquid atomic force microscopy characterization (Extended Data Fig. 8a). The fresh SLB was covered with 150 µl of 0.05 mg ml^{-1} RHP1 solution for 30 min. The RHP1–SLB solution was subjected to a liposome buffer rinse to remove free polymers. Current–voltage and cyclic voltammetry measurements were performed on a home-built system (Extended Data Fig. 8b), and the membrane permeability was fitted using a model described previously[19].

Data availability

The data supporting the findings of this study, as well as descriptions of the methodologies used in the code, are available within the article and the Extended Data items. For reproduction purposes, the raw data used to generate the figures and input scripts used to produce the simulations are available from the Dryad Digital Repository (https://doi.org/10.6078/D1VX0B).

Code availability

All custom scripts (sequence analyses and HMM) are available from the corresponding author upon request.

28. Stewart, J. C. M. Colorimetric determination of phospholipids with ammonium ferrothiocyanate. *Anal. Biochem*. **104**, 10–14 (1980).
29. Lira, R. B., Steinkuhler, J., Knorr, R. L., Dimova, R. & Riske, K. A. Posing for a picture: vesicle immobilization in agarose gel. *Sci. Rep*. **6**, 25254 (2016).
30. Manneville, J. B. et al. COPI coat assembly occurs on liquid-disordered domains and the associated membrane deformations are limited by membrane tension. *Proc. Natl Acad. Sci. USA* **105**, 16946–16951 (2008).
31. Hess, B., Kutzner, C., van der Spoel, D. & Lindahl, E. GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation. *J. Chem. Theory Comput*. **4**, 435–447 (2008).
32. Klauda, J. B. et al. Update of the CHARMM all-atom additive force field for lipids: validation on six lipid types. *J. Phys. Chem. B* **114**, 7830–7843 (2010).
33. Huang, J. et al. CHARMM36m: an improved force field for folded and intrinsically disordered proteins. *Nat. Methods* **14**, 71–73 (2017).
34. MacKerell, A. D. et al. All-atom empirical potential for molecular modeling and dynamics studies of proteins. *J. Phys. Chem. B* **102**, 3586–3616 (1998).
35. Miyamoto, S. & Kollman, P. A. Settle: an analytical version of the shake and rattle algorithm for rigid water models. *J. Comput. Chem*. **13**, 952–962 (1992).
36. Martínez, L., Andrade, R., Birgin, E. G. & Martinez, J. M. PACKMOL: a package for building initial configurations for molecular dynamics simulations. *J. Comput. Chem*. **30**, 2157–2164 (2009).
37. Parrinello, M. & Rahman, A. Polymorphic transitions in single crystals: a new molecular dynamics method. *J. Appl. Phys*. **52**, 7182–7190 (1981).
38. Darden, T., York, D. & Pedersen, L. Particle mesh ewald: an *N*·log(*N*) method for Ewald sums in large systems. *J. Chem. Phys*. **98**, 10089–10092 (1993).
39. Essmann, U. et al. A smooth particle mesh Ewald method. *J. Chem. Phys*. **103**, 8577–8593 (1995).
40. Lee, J. et al. CHARMM-GUI input generator for NAMD, GROMACS, AMBER, OpenMM, and CHARMM/OpenMM simulations using the CHARMM36 additive force field. *J. Chem. Theory Comput*. **12**, 405–413 (2016).
41. Humphrey, W., Dalke, A. & Schulten, K. VMD: visual molecular dynamics. *J. Mol. Graph. Model*. **14**, 33–38 (1996).
42. Zeidel, M. L., Ambudkar, S. V., Smith, B. L. & Agre, P. Reconstitution of functional water channels in liposomes containing purified red-cell CHIP28 protein. *Biochemistry* **31**, 7436–7440 (1992).
43. Shen, Y. X. et al. Highly permeable artificial water channels that can self-assemble into two-dimensional arrays. *Proc. Natl Acad. Sci. USA* **112**, 9810–9815 (2015).
44. Heller, W. T. et al. The suite of small-angle neutron scattering instruments at Oak Ridge National Laboratory. *J. Appl. Cryst*. **51**, 242–248 (2018).
45. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
46. Dempster, A. P., Laird, N. M. & Rubin, D. B. Maximum likelihood from incomplete data via the EM algorithm. *J. R. Stat. Soc. A* **39**, 1–22 (1977).
47. Forney, C. D. The Viterbi algorithm. *Proc. IEEE* **61**, 268–278 (1973).

Acknowledgements This work was supported by the US Department of Defense, Army Research Office, under contract W911NF-13-1-0232 and the National Science Foundation under contract DMR-183696. M.O.d.l.C. and B.Q. acknowledge support through grant DE-FG02-08ER46539 from the Department of Energy (DOE) Basic Energy Science Office and the Center for Computation and Theory of Soft Materials, as well as computational support by the Sherman Fairchild Foundation. Z. H. and M. R. acknowledge support from the Air Force Office of Sponsored Research Award FA9550-15-1-0273. A part of this research used resources at the High Flux Isotope Laboratory, a DOE Office of Science User Facility operated by the Oak Ridge National Laboratory. We acknowledge the support of the National Institute of Standards and Technology, US Department of Commerce, in providing the neutron research facilities used in this work. We thank L. He and Y. Liu for help in the SANS studies. Scattering studies at the Advanced Light Source and RHP characterization at the Molecular Foundry were supported by the Office of Science, Office of Basic Energy

Sciences of the US DOE under contract DE-AC02-05CH11231. We thank A. A. A. Smith for help in polymer syntheses; Y. W. Qian for characterizing the liposome sizes; A. Martin and E. M. López-Alfonzo for help in stopped-flow experiments. We thank CoC-NMR for help with RHP characterization; instruments at CoC-NMR are supported in part by NIH S10OD024998.

Author contributions T.J. made the first experimental observation of RHP-based proton transport. T.X. directed the project development. T.J. and T.X. designed the experiments and data analysis. T.J. performed imaging, FRAP and transport studies on liposomes. A.H. synthesized and characterized RHPs. M.E. and Z.R. performed the RHP segmental analysis. Z.H. and M.R. performed transport studies on SLBs. B.Q. and M.O.d.l.C. performed all-atom simulations and data analysis. Y.Z. and H.H performed HMM studies. T.J. and A.D.C. prepared and characterized liposomes. T.J. and W.T.H. collected and analysed SANS data with T.X. All authors participated in the writing of the manuscript.

Competing interests The authors declare no competing interests.

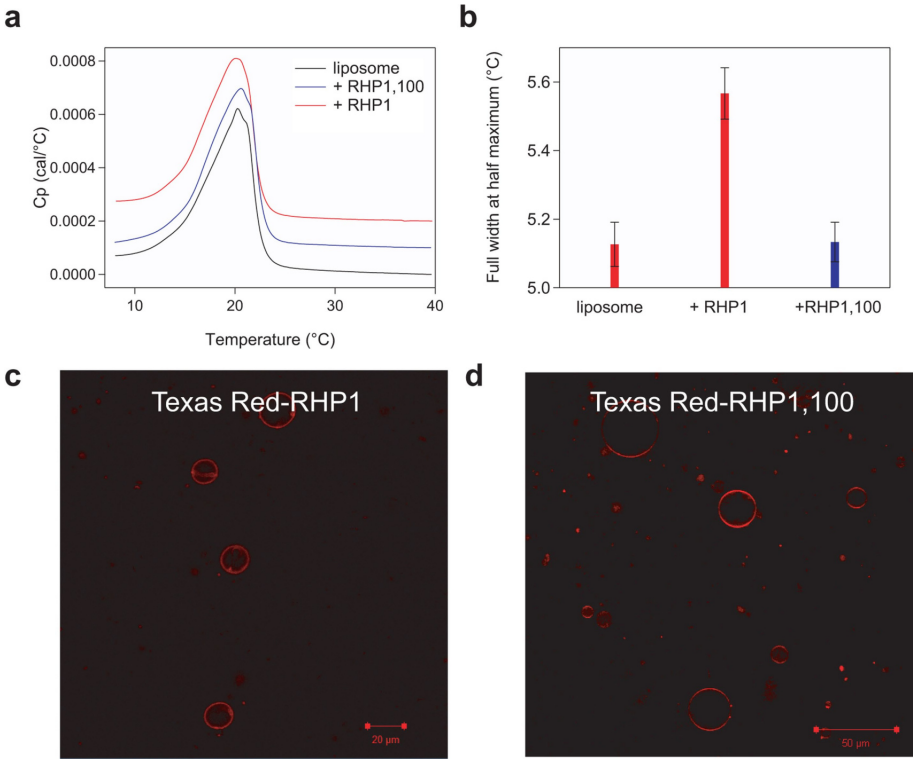
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1881-0.

Correspondence and requests for materials should be addressed to T.X.

Peer review information *Nature* thanks Junli Hou, Kenichi Kuroda and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

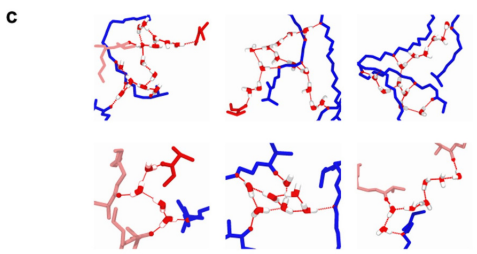
Reprints and permissions information is available at http://www.nature.com/reprints.

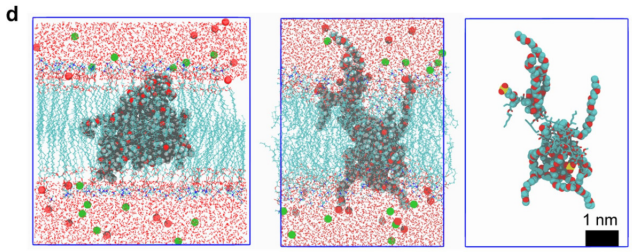


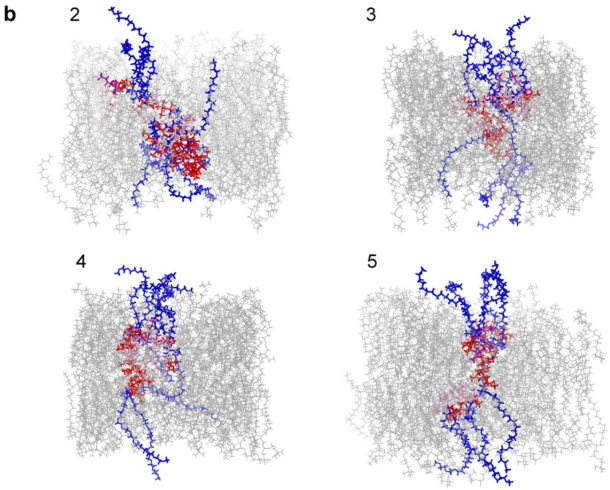
Extended Data Fig. 1 | Interactions between RHPs and liposomes. a, DSC profiles of LUVs and RHP1- and RHP1,100-containing LUVs (POPE:POPG molar ratio, 3:1). LUVs were grown with RHP1 or RHP1,100 with a lipid-to-RHP molar ratio of 50:1 and a lipid concentration of 1 mg mL^{-1}. The background signals of the buffer and the polymers were subtracted. No phase transition was observed for RHP itself. The plots are shifted vertically for clarity by +0.0001 for RHP1,100 and +0.0002 for RHP1. **b**, Full-width at half-maximum of the lipid phase transition peaks shown in **a**. Error bars are 1 s.d. ($n = 3$). **c**, **d**, Representative confocal fluorescence image of POPC GUVs incubated with Texas Red-labelled RHP1 (**c**) and Texas Red-labelled RHP1,100 (**d**).

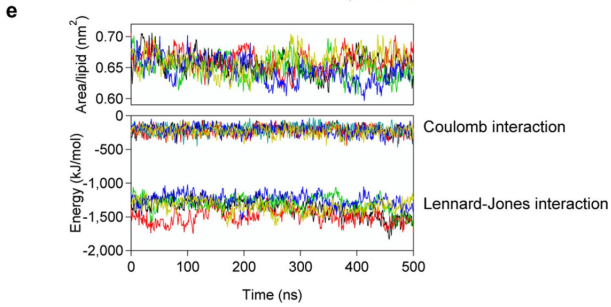
a

RHP1 chains	monomer sequences
1	11114114211231121111212131111323323322
2	31412212331123322213121213311111111112114
3	13121234123233143112112111311121311112222
4	13231222313113311111221312111113121212244
5	31111111211212123422114111123121123233333
6	41332213233212222121111131211311111114311
7	31321231221121322242321143311311111111111
8	12113111232311134121321222112331121141113
9	11121112211311113122311112212312312343334

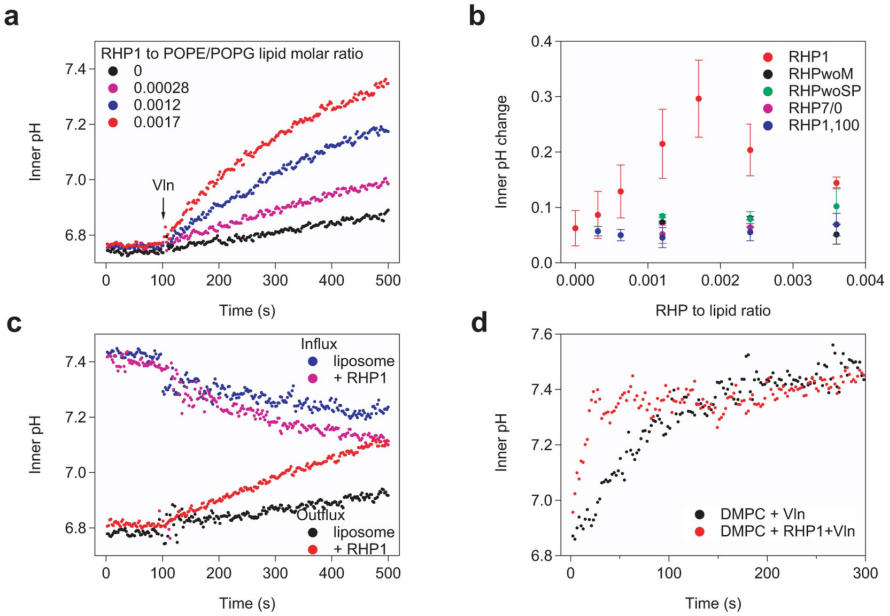




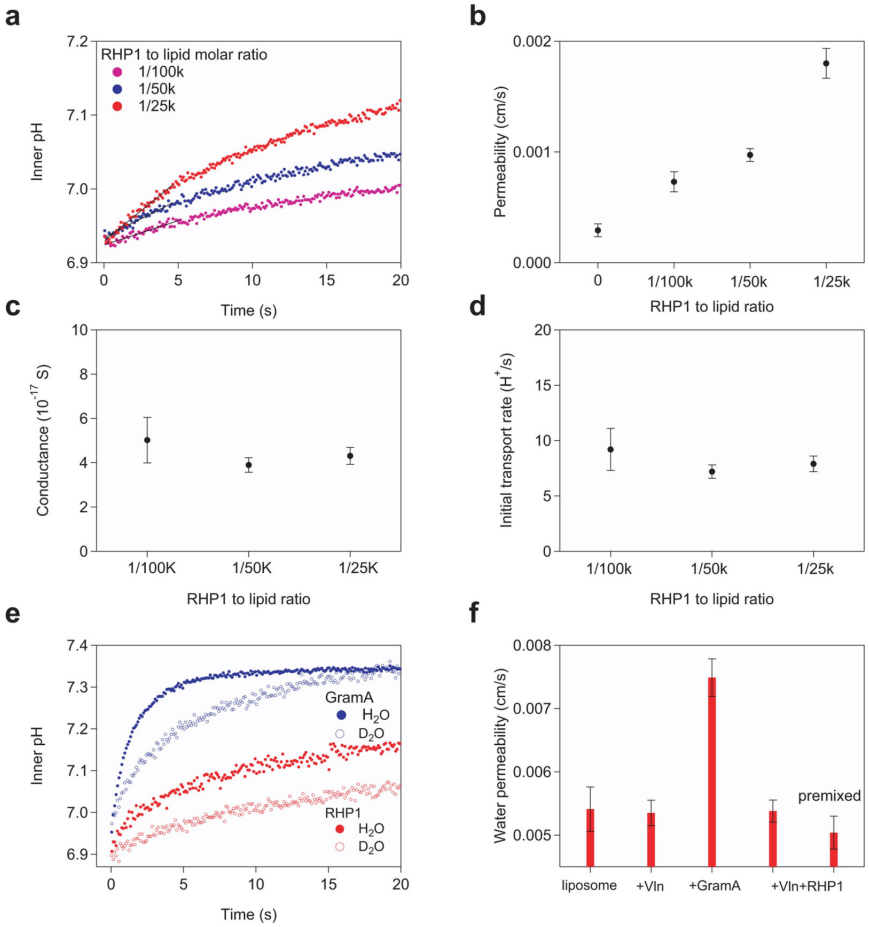




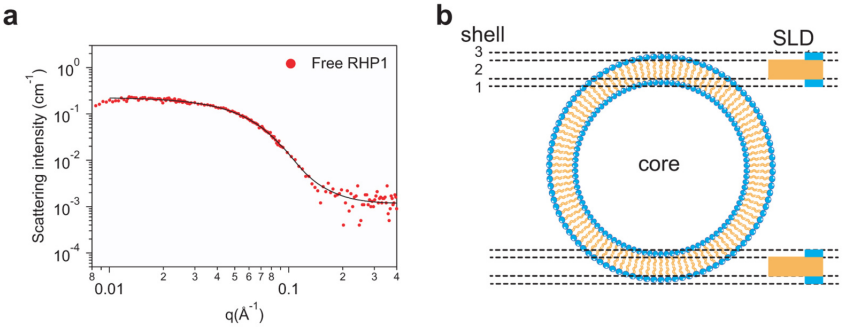
Extended Data Fig. 2 | All-atom molecular dynamics simulations. **a**, Sequences of the nine RHP1 chains used in the simulations. **b**, Snapshot of the 2nd–5th RHP1 chains in the POPC lipid bilayer in the simulations. MMA, EHMA, OEGMA and SPMA are coloured red, pink, blue and purple, respectively. Lipids are shown in grey colour. **c**, Representative HBCs in the transmembrane regions of RHP1s. The oxygen and hydrogen atoms of water molecules are coloured red and white, respectively. Hydrogen bonds are shown with red dashed lines. Lipids are omitted for clarity. **d**, Snapshots of the initial (left) and final (right) RHP1 conformations in the simulations. The RHP1 chain is highlighted, along with K^+ (orange) and Cl^- (green) ions. **e**, Convergence of the simulations. Top, area per lipid as a function of the simulation time. Bottom, Coulomb and Lennard–Jones interactions between POPC lipids and the 1st–5th RHP1s as a function of simulation time. Different colours denote the results from five parallel molecular dynamics simulations.



Extended Data Fig. 3 | Transmembrane proton transport. a, pH gradient-driven proton flux through RHP1 in the LUVs of POPE/POPG (molar ratio 3:1). **b**, Inner pH changes in POPC LUVs incubated with RHP1 and RHP variants over 200 s after adding Vln. 'RHPwoM', 'RHPwoSP' and 'RHP7/0' represent the RHP without MMA, the RHP without SPMA and the RHP without EHMA, respectively. **c**, pH gradient-driven proton flux in the influx and outflux directions. RHP1 was added to the solutions of preformed LUVs. **d**, pH gradient-driven proton flux through RHP1 in DMPC SUVs at 37 °C. RHP1-containing SUVs were prepared as indicated in the Methods section 'SANS investigation' using H_2O-based buffers, with an RHP1:DMPC ratio of 0.001.



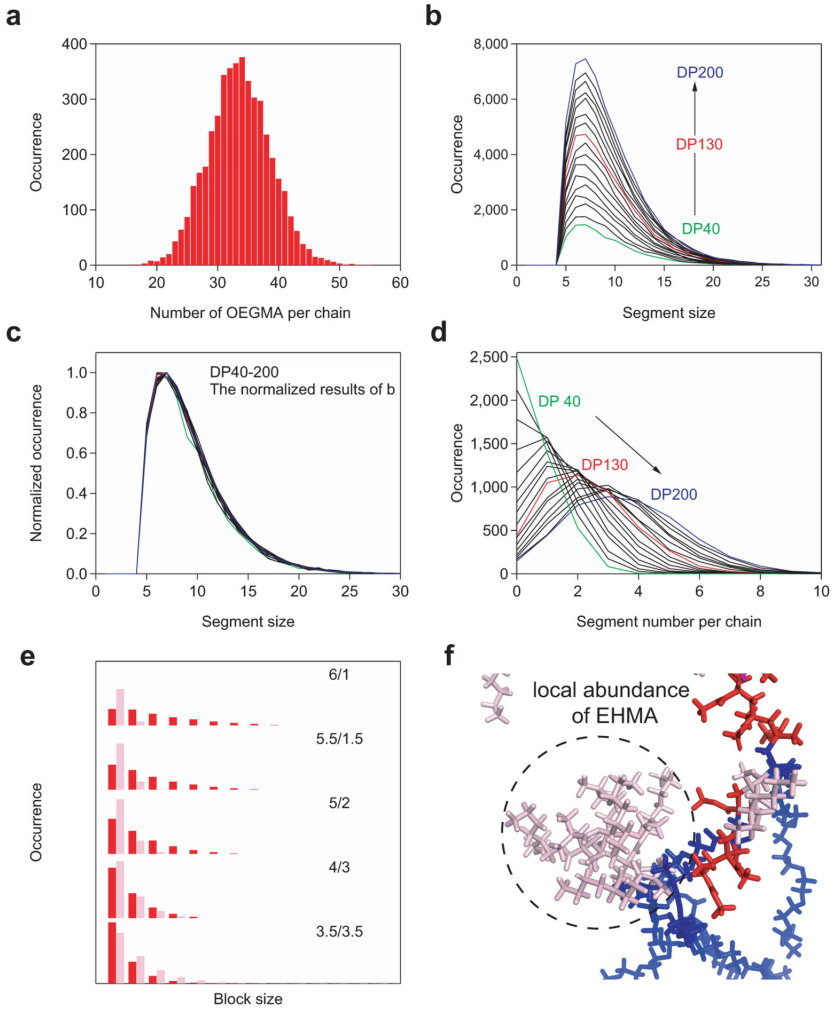
Extended Data Fig. 4 | Proton and water transport rates. a, Representative stopped-flow fluorescence traces of RHP1-containing POPC LUVs after rapid exposure to a buffer of a higher pH value. The initial and final equilibrated pH values are 6.85 and 7.35, respectively. Vln was added before the measurements. **b**, Calculated permeability values of RHP1-containing LUVs at various RHP1-to-lipid ratios. Error bars are 1 s.d. ($n = 3$). **c**, **d**, Conductance (**c**) and initial proton transport rate (**d**) per RHP1 chain at various RHP1-to-lipid ratios. Error bars are 1 s.d. ($n = 3$). **e**, Stopped-flow fluorescence traces of RHP1-containing LUVs after rapid exposure to a buffer of a higher pH or pD value. The initial and final equilibrated pH values are 6.85 and 7.35, respectively, for the H_2O-based assay. The initial and final pD values are 7.25 and 7.75, respectively, for the D_2O-based assay. Vln was added to the RHP1-containing LUV solutions, except for the GramA-containing LUVs. **f**, Osmotic water permeability of Vln, GramA and RHP1+Vln in POPC LUVs at 7 °C. Vln (K^+ channel) and GramA (proton/cations/water channel) are the negative and positive controls, respectively. GramA, Vln and RHP1 were added to the solutions of preformed LUVs. 'Premixed' denotes premixing of RHP1 and the lipid before the preparation of LUVs. Error bars are 1 s.d. ($n = 7$).



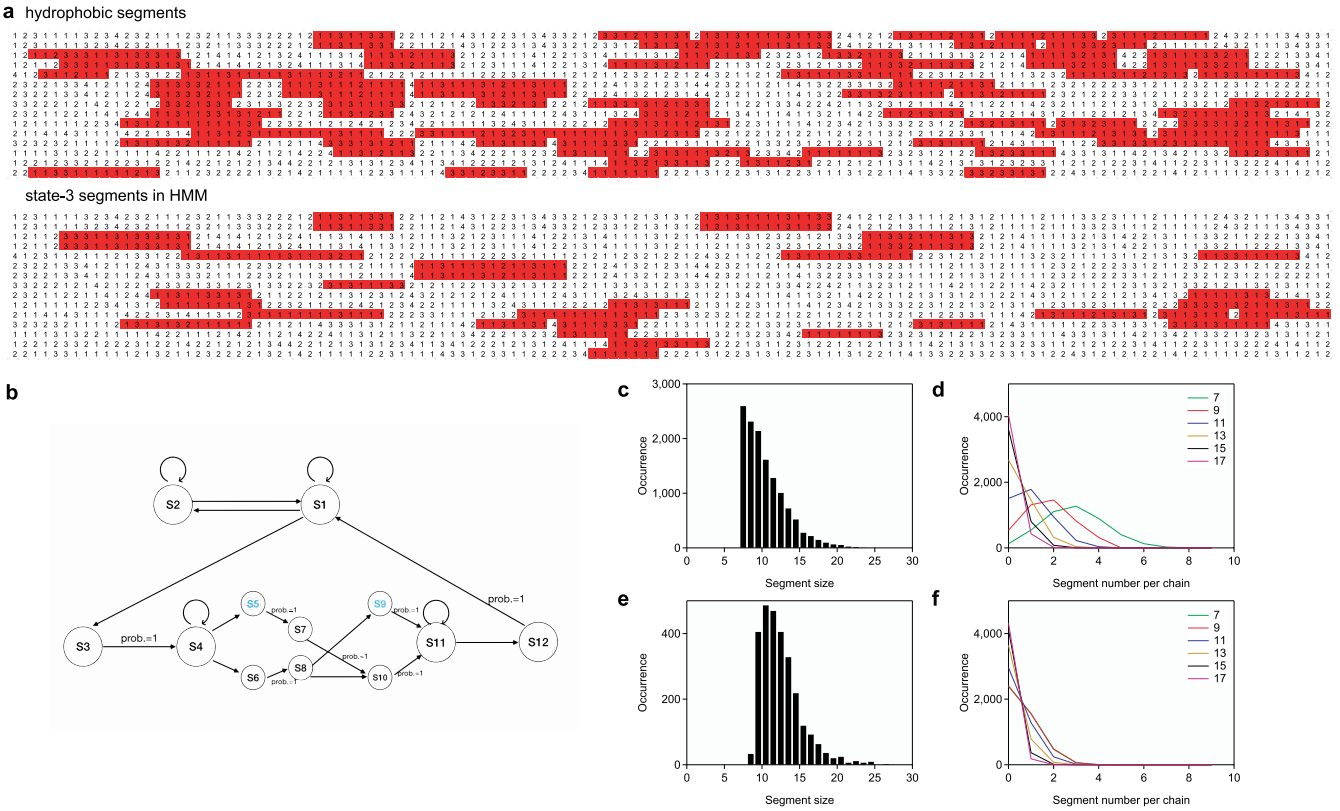
c

	liposome only	RHP1 only	RHP1/lipid 0.0005	0.001	0.002
liposome core radius (Å)	500f		500f	500f	500f
shell 1 thickness (Å)	9.32		9.92	12.1	10.5
shell 2 thickness (Å)	29.3		28.3	27.4	28.4
shell 3 thickness (Å)	7.24		6.35	3.85	5.70
RHP1 radius (Å)		23.3	17.4	26.8	22.5
RMS deviation of RHP1 radius (Å)		9.79	8.87	14.2	11.0

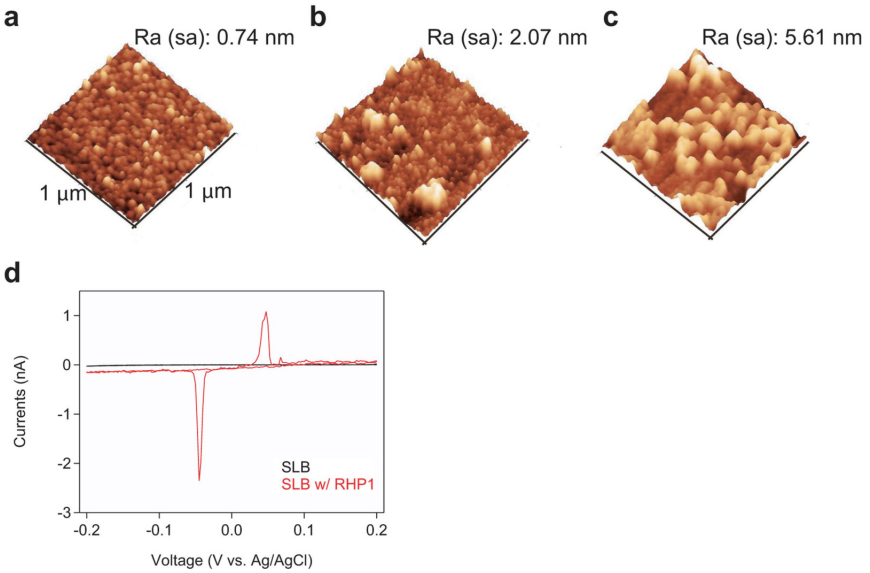
Extended Data Fig. 5 | SANS investigations. a, SANS profiles of free RHP1 in 86% D_2O. The solid line denotes a fit using a polydisperse sphere model for RHP1. The signal of the solvent has been subtracted in the presented curves. **b**, Schematic illustration of the three-shell vesicle model for the SUV, showing the regions of the liposome core (core), inner bilayer headgroup (shell 1), lipid bilayer core (shell 2) and outer bilayer headgroup (shell 3). **c**, Structural parameters obtained by fitting the SANS data with the three-shell vesicle model for the liposome and the polydisperse sphere model for RHP1. Parameters that were fixed during the fitting are denoted with 'f'; RMS represents root mean square. SLDs that were fixed were determined from the known sample composition, including when RHP1 was present in the bilayer, because the concentration was too low to impact the SLD value of shells significantly. The SLD values used for the liposome core, for shells 1, 2, 3 and for the solvent are 5.41, 1.88, 6.84, 1.88 and 5.41×10^{-6} Å^{-2}, respectively. The SLD value for RHP1 is 1.4×10^{-6} Å^{-2} in the sphere model.



Extended Data Fig. 6 | Sequence analyses of RHPs. a, Statistical distribution of number of OEGMA monomers per RHP1 chains in the 4,500 simulated RHP1 sequences (DP = 130). This distribution reflects the compositional dispersity in RHP1 chains. **b**, Statistical size distribution of hydrophobic segments with one OEGMA in RHP1s of various chain lengths ranging from DP = 40 to 200. The number of chains (4,500) was kept constant for each DP analysis. **c**, Normalized profiles of the size distribution in **b**. **d**, Number of hydrophobic segments per chain at various chain lengths. The counted hydrophobic segments here contain one OEGMA with segment size ≥11. **e**, Statistical size distributions of blocks of EHMA (pink) and MMA (red) within hydrophobic segments containing one OEGMA in RHP variants. Results are shown for hydrophobic segments with segment size ≥11. **f**, Schematic illustration of five EHMA monomers distributed at the end of an inserted RHP1 chain in a molecular dynamics simulation. EHMA, MMA and OEGMA are coloured pink, red and blue, respectively. Lipids are omitted for clarity.



Extended Data Fig. 7 | HMM prediction. a, Comparison between state-3 segments (size ≥7; top) assigned using the method described in Methods section 'Hydrophobic segments' and those predicted by HMM (bottom). '1', '2', '3' and '4' denote MMA, OEGMA, EHMA and SPMA, respectively. **b**, Transition diagram with a state-3 segment of size ≥7. States 1 and 2 are denoted as S1 and S2, respectively. State 3 here is broken into 10 states to comply with the rules outlined in Methods section 'HMM prediction'. S5 and S9 are the special states in a state-3 segment that must emit OEGMA with probability 1 (denoted as 'prob. = 1'). **c**, Statistical size distribution of the state-3 segments predicted by HMM in 4,500 RHP1 sequences. **d**, Number of the predicted state-3 segments per polymer chain. **e**, Statistical size distribution of predicted state-3 segments containing one OEGMA. **f**, Number of predicted state-3 segments containing one OEGMA per polymer chain. 4,375 out of 4,500 RHP1 chains contain at least one state-3 segment.



Extended Data Fig. 8 | Proton transport on SLBs. a, Liquid atomic force microscopy images of a Pd surface covered with SLBs, showing a surface roughness of about 0.74 nm. **b**, Image for SLB+RHP1 (RHP1-to-lipid molar ratio, 0.001), showing a roughness of about 2.07 nm. **c**, Surface covered with SLB+RHP1 (RHP1-to-lipid molar ratio, 0.01), showing a roughness of about 5.61 nm. **d**, Current–voltage sweep of the protonic device. The Pd/SLB contact has a current of <0.01 nA across the entire voltage range, confirming the high polarization resistance of the SLB (black trace). With RHP1 incorporated, an oxidation peak is observed at about 48 mV, indicating H^+ flow across the SLB; a reduction peak is observed at about −45 mV, indicating conduction of H across the SLB.

Extended Data Table 1 | Physical-chemical properties of RHPs

	MMA	EHMA	OEGMA	SPMA	Mn(kg/mol)*	Mn(kg/mol)†	Đ†
RHP1	50 (50.8)	20 (21.5)	25 (24.1)	5 (3.7)	30.6	29.0	1.3
RHP70+0	70 (72.1)	0 (0)	25 (23.1)	5 (4.8)	18.0	20.2	1.4
RHP60+10	60 (61.9)	10 (10.0)	25 (23.5)	5 (4.6)	18.5	21.0	1.3
RHP55+15	55 (51.9)	15 (12.2)	25 (33.3)	5 (2.5)	28.0	33.6	1.6
RHP40+30	40 (40.4)	30 (30.3)	25 (24.3)	5 (5.0)	26.5	22.0	1.3
RHP35+35	35 (36.1)	35 (36.4)	25 (23.8)	5 (3.8)	40.0	51.1	1.3
RHPwoSP	53	21	26	0	29.2	44.4	1.3
RHPwoM	0	40	50	10	35.6	45.9	1.2
RHP1-7k	50	20	25	5	7.0	16.3	1.2
RHP1-19k	50	20	25	5	18.7	18.8	1.4
RHP1-40k	50	20	25	5	40.6	40.2	1.3
RHP1,100	50	20	25	5		57.8	1.2

*Determined by NMR.

†Measured by gel permeation chromatography, using PEG standards for calibration.

‡Numbers in parentheses are the monomer compositions determined by ^{13}C NMR (150 MHz).

Article

Palaeoclimate evidence of vulnerable permafrost during times of low sea ice

https://doi.org/10.1038/s41586-019-1880-1

Received: 23 December 2018

Accepted: 24 October 2019

Published online: 8 January 2020

A. Vaks[1,2]*, A. J. Mason[2], S. F. M. Breitenbach[3], A. M. Kononov[4], A. V. Osinzev[5], M. Rosensaft[1], A. Borshevsky[1], O. S. Gutareva[4] & G. M. Henderson[2]

Climate change in the Arctic is occurring rapidly, and projections suggest the complete loss of summer sea ice by the middle of this century[1]. The sensitivity of permanently frozen ground (permafrost) in the Northern Hemisphere to warming is less clear, and its long-term trends are harder to monitor than those of sea ice. Here we use palaeoclimate data to show that Siberian permafrost is robust to warming when Arctic sea ice is present, but vulnerable when it is absent. Uranium–lead chronology of carbonate deposits (speleothems) in a Siberian cave located at the southern edge of continuous permafrost reveals periods in which the overlying ground was not permanently frozen. The speleothem record starts 1.5 million years ago (Ma), a time when greater equator-to-pole heat transport led to a warmer Northern Hemisphere[2]. The growth of the speleothems indicates that permafrost at the cave site was absent at that time, becoming more frequent from about 1.35 Ma, as the Northern Hemisphere cooled, and permanent after about 0.4 Ma. This history mirrors that of year-round sea ice in the Arctic Ocean, which was largely absent before about 0.4 Ma (ref. [3]), but continuously present since that date. The robustness of permafrost when sea ice is present, as well as the increased permafrost vulnerability when sea ice is absent, can be explained by changes in both heat and moisture transport. Reduced sea ice may contribute to warming of Arctic air[4–6], which can lead to warming far inland[7]. Open Arctic waters also increase the source of moisture and increase autumn snowfall over Siberia, insulating the ground from low winter temperatures[8–10]. These processes explain the relationship between an ice-free Arctic and permafrost thawing before 0.4 Ma. If these processes continue during modern climate change, future loss of summer Arctic sea ice will accelerate the thawing of Siberian permafrost.

Arctic Ocean sea ice has declined increasingly rapidly in recent decades, with progressive ice thinning and increasing areas of open water during the summer[11]. Complete loss of summer sea ice is expected by the mid-twenty-first century[1]. The recent loss of Arctic sea ice raises concerns about its effects on other aspects of the global climate system, including potential acceleration of permafrost thawing[7]. Permafrost degradation as a result of anthropogenic global warming could amplify climate change by releasing large volumes of carbon stored in permafrost in the form of carbon dioxide and methane[12]. In addition, permafrost thawing increases thermokarst development, coastal erosion and liquefaction of ground previously cemented by ground ice, endangering infrastructures relying on permafrost as solid ground[13]. Establishing the relationship between loss of Arctic sea ice and permafrost response is therefore an important goal.

Understanding of the response of permafrost to changing climate can be improved with knowledge of past environmental conditions. Precisely dated growth periods of speleothems (stalagmites, stalactites and flowstones) from caves located in permafrost regions have proved an effective tool for the reconstruction of past permafrost extent and continuity[14]. Speleothems grow only when meteoric waters seep through the vadose zone into caves. When the temperature in the upper vadose zone falls below 0 °C throughout the year, water freezes, infiltration stops and speleothem growth ceases. Speleothems found in modern permafrost regions are thus relicts from warmer periods when permafrost was absent, or when permafrost thawed temporarily[14,15]. Dating of these relicts allows a comparison of periods of permafrost absence to other aspects of the past environment.

Here, we reconstruct permafrost dynamics in central Eastern Siberia over the past 1,500 thousand years, using uranium–lead (U–Pb)-dated speleothems from Ledyanaya Lenskaya Cave (60° 22′ 15.60″ N–116° 56′ 47.30″ E) (Fig. 1) located in the zone of present-day continuous permafrost (that is, year-round frozen ground across the whole region[16]; Extended Data Fig. 3). The study area is characterized by cold continental climate (Dfc according to the Köppen classification[17]), with mean July and January air temperatures of +18 °C and −32 °C, respectively, a mean annual air temperature (MAAT) of about −6 °C (Extended Data Fig. 2), and annual precipitation of about 400 mm. The record from Ledyanaya Lenskaya Cave is compared with data from Botovskaya Cave

[1]Geological Survey of Israel, Jerusalem, Israel. [2]Department of Earth Sciences, University of Oxford, Oxford, UK. [3]Department of Geography and Environmental Sciences, Northumbria University, Newcastle upon Tyne, UK. [4]Institute of Earth's Crust, Russian Academy of Sciences, Siberian Branch, Irkutsk, Russia. [5]Speleoclub Arabica, Irkutsk, Russia. *e-mail: antonv@gsi.gov.il

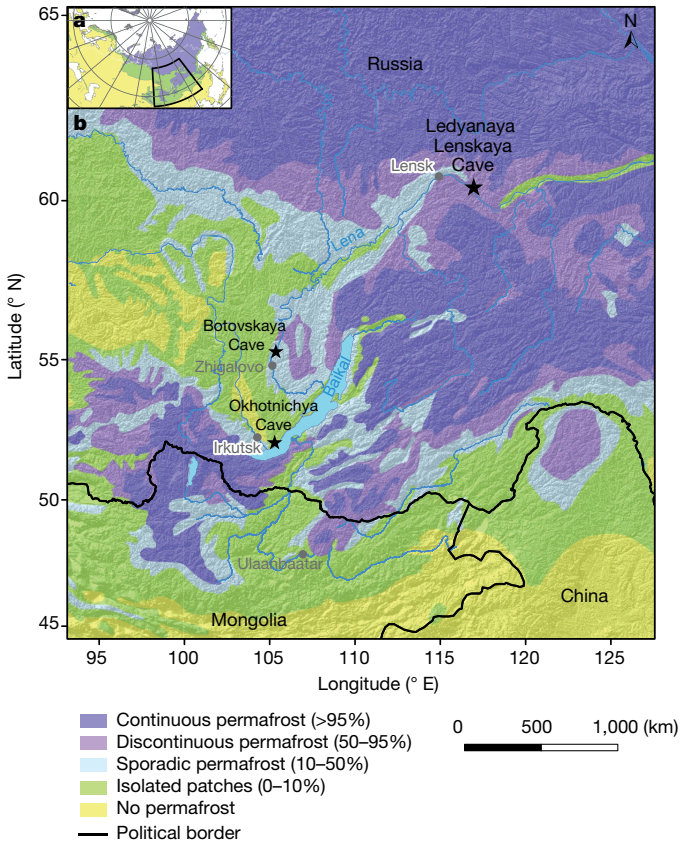


Fig. 1 | Study area and permafrost maps. a, Permafrost map of northern Eurasia with the study area marked by the black rectangle, continuous permafrost in purple, discontinuous permafrost in green and the area with no permafrost in yellow. **b**, Extent of permafrost types in eastern Siberia[16] and the location of Ledyanaya Lenskaya, Botovskaya and Okhotnichya caves, marked by black stars. Cities and towns are marked by grey circles. Permafrost types (see legend) are defined by the percentage of the year-round frozen ground.

(55° 17′ 59.03″ N–105° 19′ 46.02″ E) (Fig. 1), located in an area of discontinuous permafrost with a MAAT of about −2 °C (Extended Data Figs. 2, 3).

This study continues the research of Vaks et al.[14], who found that, in Ledyanaya Lenskaya Cave, the most recent permafrost thaw occurred at 429 ± 23 thousand years ago (ka) (the original age cited by Vaks et al.[14] is 427 ± 23 ka; the age above is re-calculated using updated half-lives of ^{234}U and ^{230}Th (ref. [18])) during the warmth of Marine Isotope Stage (MIS) 11[19]. That study was limited, however, by the approximately 500 ka range of uranium–thorium (U–Th) chronology, meaning that older samples could not be analysed. U–Pb chronology enables dating of such older speleothems. Here, we use 52 U–Pb ages of 11 speleothems from Ledyanaya Lenskaya Cave to extend the age range of known Siberian permafrost history. A smaller number of ages were also determined for three samples from the more southerly and warmer Botovskaya Cave[14] (Fig. 1) (also see Methods, Extended Data Figs. 1, 3–9, and Supplementary Tables 1 and 2 for chronological methods and full results). Ages indicate a division of speleothem deposition in Ledyanaya Lenskaya Cave, and therefore of permafrost presence, into three distinct periods (Figs. 2a, 3a).

During the period from 1,500 to about 1,350 ka, speleothems apparently grew continuously (within the limits of analytical uncertainties), suggesting discontinuous or absent permafrost above the cave. Globally, this interval spans MIS 50 to MIS 43 and is characterized by glacial–interglacial cycles with 41,000-yr periodicity[19]. Most of the analysed speleothems in Ledyanaya Lenskaya Cave grew in this period (7 out of 11) (Extended Data Fig. 4). These oldest vadose speleothems were the first deposited directly on the cave host rock, heralding the onset of vadose conditions at this site, and suggesting that, before 1,500 ka, the cave may have been located below the local groundwater level. The current groundwater table is located about 50 m below the cave's entrance and controlled by the nearby Lena River.

The period from about 1,350 ka to about 400 ka is defined by intermittent speleothem growth with long-lasting hiatuses without speleothem deposition. These growth cessations probably indicate continuous periods of permafrost and are found at the time of most glacial MISs and some interglacial MISs. Speleothem growth, demonstrating the absence of permafrost, occurred during most interglacials (Fig. 2a). From about 1,300 ka, speleothems only grew in the shallower portion of the cave (15–20 m below the surface) and not in the deeper area (about 60 m) (Figs. 2a, 3a, Extended Data Fig. 1). This may indicate that the permafrost was thawing only to depth of 15–20 m, whereas relict permafrost remained at greater depth, showing that the duration of thawing periods was relatively short and/or that MAATs were reduced compared with the period before 1,300 ka.

From about 400 ka until present, speleothem growth ceased completely and permafrost appears to have been continuous above Ledyanaya Lenskaya Cave[14] (Figs. 2a, 3a).

Caves located farther south near Lake Baikal (Botovskaya and Okhotnichya) (Fig. 1) show speleothem deposition during warm periods during the past 700 ka (this study and Vaks et al.[14] (Fig. 2a)). These southerly caves indicate that climate in southern Siberia was warmer than in Ledyanaya Lenskaya Cave, enabling deposition of speleothems, while in Ledyanaya Lenskaya Cave to the north, speleothem growth ceased completely for the entire past ~400 ka.

On the basis of data for the past 500 ka, Vaks et al.[14] found permafrost thawing at Ledyanaya Lenskaya Cave during the unusual warmth of MIS 11 (429 ± 23 ka) but not in younger interglacials. They suggested that an increase in global mean surface temperature of 1.5 °C (above pre-industrial levels) represents the threshold above which continuous permafrost thaws at its southern fringes. Our results indicate that substantial speleothem deposition occurred before MIS 11, when global mean surface temperatures (as indicated by Pacific warm pool sea surface temperatures (SSTs)) were lower than those of MIS 11 (for example, MISs 25, 19 and 15) and even lower than today (for example, MIS 23; Fig. 2c)[20]. These earlier periods of speleothem growth indicate that global mean surface temperature is not the only control on the extent of Siberian permafrost. Other possible controls may include: (1) local summer insolation; (2) palaeogeographic changes; (3) greater poleward heat transport in the Northern Hemisphere, leading to relatively warmer conditions in the North Atlantic, Arctic, and/or over the Eurasian landmass; or (4) Arctic Ocean SSTs and the extent of Arctic summer sea-ice cover.

The intensity of summer insolation at latitude 60° N (Fig. 2d)[21] may directly affect Siberian summer temperatures and therefore influence permafrost thawing. Many periods of speleothem deposition occurred when July insolation was high (that is, >500 W m^{-2}), but there is no direct relationship between insolation and thawing of permafrost. No thawing took place during periods of insolation >500 W m^{-2} during the past 400 ka, but thawing did occur at much lower insolation earlier in the record. Local summer insolation is therefore not the key factor determining the presence of permafrost above Ledyanaya Lenskaya Cave.

Arctic palaeogeographic conditions during interglacials of the entire past 1,500 ka were similar to present, with the open Bering Strait enabling water exchange between the Pacific and Arctic oceans[22]. The Atlantic meridional overturning circulation (AMOC) transports heat from the tropics to the northern Atlantic Ocean, thereby increasing the heat flux to high latitudes in continental Eurasia and the Arctic[23]. The period between about 2.4 and 1.3 million years ago (Ma) was characterized by enhanced AMOC, causing heat piracy from the Southern Hemisphere to the Northern Hemisphere, which was consequently

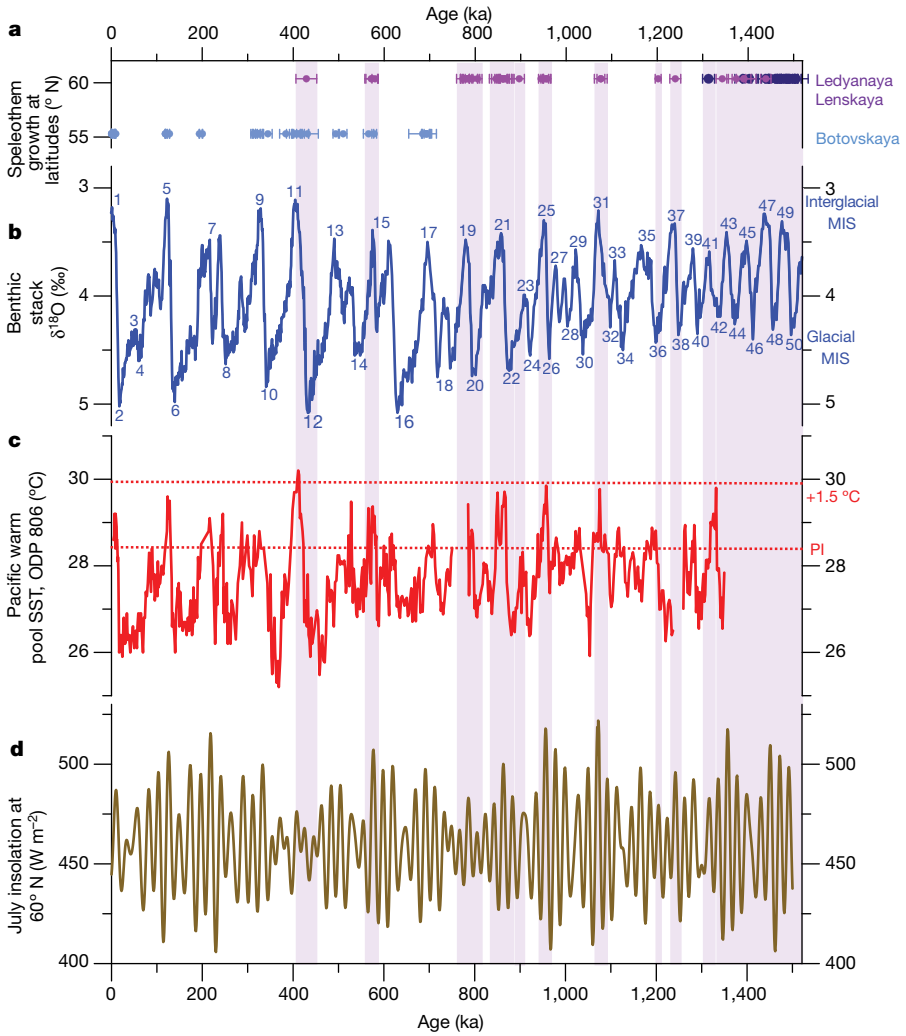


Fig. 2 | Siberian speleothem deposition periods compared with records of MISs, Pacific warm pool SSTs and July mean insolation at 60° N. **a**, Distribution of speleothem U–Pb and U–Th ages (±2σ) in Ledyanaya Lenskaya Cave (purple circles) and in Botovskaya Cave (light blue circles) in time (ka) and space (latitude ° N). Ages of Ledyanaya Lenskaya SLL14 speleothems (60 m below the surface) are marked by dark purple circles, and of SLL9/SLL10 speleothems (15–20 m below the surface) by light purple circles. Purple shaded areas show how periods of speleothem growth in Ledyanaya Lenskaya Cave relate to other climatic records. **b**, Benthic $\delta^{18}O$ stack[19] with glacial MIS numbers below and interglacial MIS numbers above, **c**, Pacific warm pool magnesium/calcium-inferred SST changes during the past 1,400 ka, which are considered to be a reasonable reflection of changes in the mean Earth surface temperature[20]. The pre-industrial SST is marked by the lower dotted red horizontal line and the abbreviation 'PI'. The SST 1.5 °C higher than the pre-industrial level is marked by the upper dotted red line and the abbreviation '+1.5 °C', Ocean Drilling Program (ODP) site 806. **d**, July mean insolation at 60° N[21].

relatively warm[2]. This enhanced northward heat flux caused substantially higher SSTs in the North Atlantic than those of most of the Middle-Late Quaternary and Holocene[2] (Fig. 3c, d). After about 1,300 ka, the AMOC gradually weakened, leading to a concurrent decrease of the North Atlantic SST of 1.5 °C to 3 °C (ref. [24]; Fig. 3c, d). This is likely to cause progressive cooling of the Eurasian landmass on a long-term scale and could influence the presence of permafrost (which is much less common early in the record when the Northern Hemisphere was warmer on average). Again, there is no simple relationship between speleothem growth in Ledyanaya Lenskaya Cave and North Atlantic warmth: periods of significant North Atlantic warmth during the past 400 ka are not associated with permafrost thawing. For example: MIS 9 (Fig. 3c, d) and MIS 5 (Fig. 3d) were warmer than all other times in the past about 1,300 ka, but there is no permafrost thawing above Ledyanaya Lenskaya Cave. Of these, MIS 5 is particularly notable because both the Pacific warm pool (Fig. 2c) and North Atlantic (Fig. 3d) were warmer than many earlier thaw periods, and as warm as interglacials in the period 1,500–1,300 ka when the permafrost above the Ledyanaya Lenskaya Cave was discontinuous or absent. Yet, there is no evidence for permafrost thawing above Ledyanaya Lenskaya Cave at MIS 5 (Fig. 3a).

It is striking that the permanent presence of permafrost since 400 ka initiates at the same time as perennial sea ice is established in the Arctic Ocean (Fig. 3a, e). The appearance of perennial sea ice is marked by an abrupt increase in the appearance of sea-ice-associated fauna in the western Arctic Ocean[3] (Fig. 3e) and the disappearance from the Arctic Ocean of fauna that today is found in the North Atlantic[25]. The perennial Arctic sea ice (Fig. 3e) remained intact even during MIS 9 and MIS 5e, when both the tropics (Fig. 2c) and North Atlantic (Fig. 3c, d) were particularly warm.

Climatic models for present and past climates indicate a relationship between Arctic sea ice and Eurasian permafrost caused by changes in atmospheric heat transport[5,26]. Removal of Arctic sea ice warms the air above the sea surface[5], increasing the moisture content of the atmosphere, and therefore increasing the transport of both sensible and latent heat from the ocean via atmospheric transport

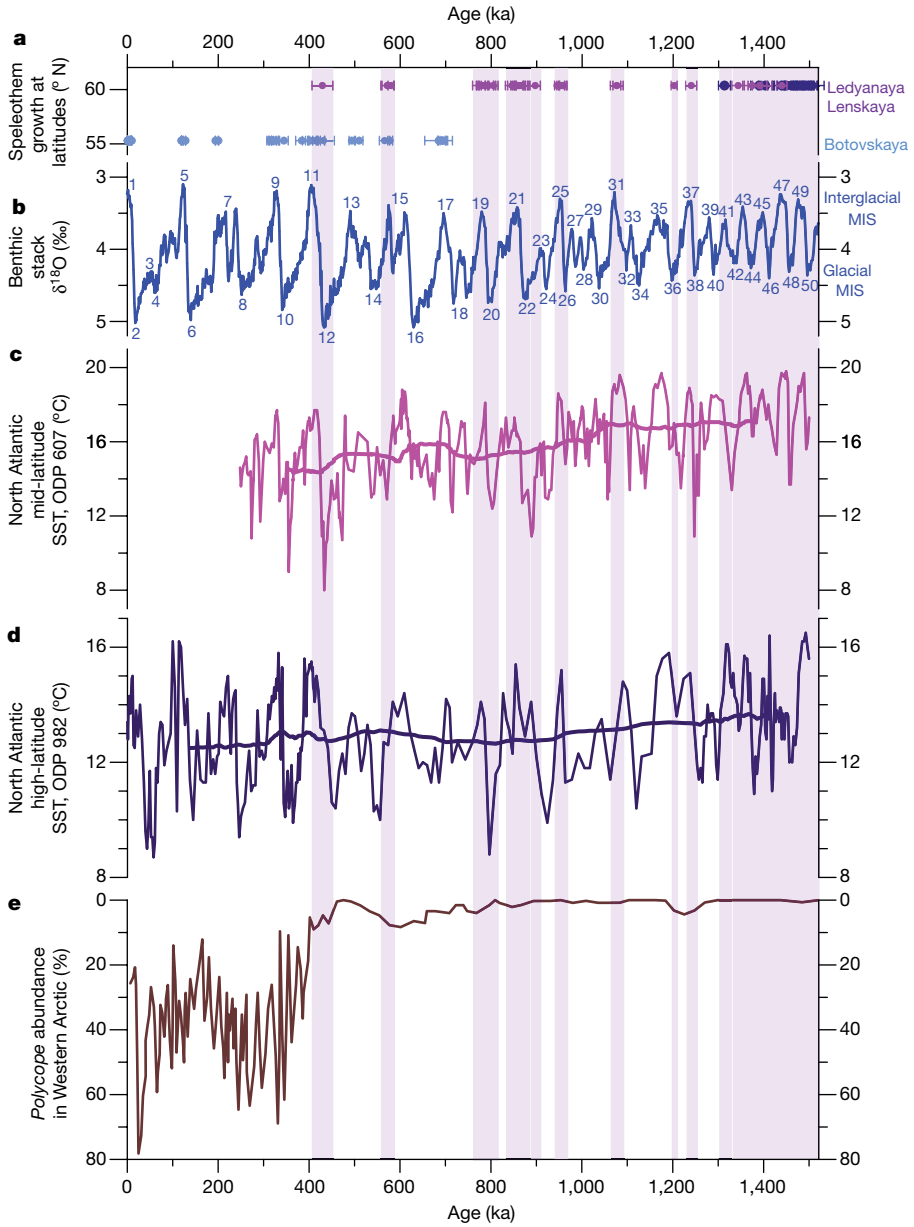


Fig. 3 | Siberian speleothem deposition periods compared with records of MISs, North Atlantic SST and presence of sea ice in the Arctic Ocean. **a**, Distribution of speleothem U–Pb and U–Th ages (±2σ) in time and space (details in caption of Fig. 2). **b**, Benthic δ^{18}O stack[19] with glacial MIS numbers below and interglacial MIS numbers above. **c**, North Atlantic mid-latitude (41° N, 33° W) alkenone unsaturation ratio SST as recorded in ODP 607 core[24], and with an 81-data-point (~200 ka) running average showing an approximately 3 °C decrease in the long-term SST. **d**, North Atlantic high-latitude (58° N, 16° W) alkenone unsaturation ratio SST as recorded in ODP 982 core[24], and with an 81-data-point (~200 ka) running average showing an approximately 1.5 °C decrease in the long-term SST. **e**, Per cent abundance of the genus *Polycope*, a benthic opportunistic genus signifying high local surface productivity in Arctic sea-ice margin environments, and therefore presence of the sea ice in Northwind Ridge, western Arctic Ocean[3].

to continental interiors[4]. Resulting warming can penetrate up to 1,500 km inland, peaking in autumn, leading to permafrost degradation[7]. An open Arctic Ocean also leads directly to increased transport of moisture from the sea to the continent. High deuterium-excess values of autumn atmospheric precipitation in Siberia show that the open Arctic Ocean comprises a substantial moisture source that is shut down when the Arctic ice cover is established in early winter[27]. Models show that future increase in Arctic precipitation will come mainly from evaporation from the Arctic Ocean due to retreating sea ice, and not from enhanced moisture transport from lower latitudes[28]. At present, decreased sea-ice cover in the autumn increases moisture and leads to heavier autumn snowfall over Siberia[8,9]. Thicker autumn snow cover insulates the ground from cold winter temperatures, increasing winter and mean annual ground temperatures[9,10]. This effect, well known to influence seasonal vegetation[10], has not previously been recognized as a substantial long-term control on the extent of permafrost, but the strong relationship between perennial sea ice and permafrost observed in this study suggests that it may be an important controlling factor. The appearance of perennial sea ice about 400 ka decreases the Arctic heat (sensible and latent) and moisture source[27], cooling the Arctic air and reducing snowfall on the continent. This may lead to poorer ground insulation and a decrease in ground temperature, assisting the establishment of continuous permafrost[29]. Future loss of summer sea ice may have the opposite effect—increasing ground temperatures and, subsequently, increasing speed of permafrost thawing.

The stability of continuous permafrost near its modern southern boundary in Siberia hinges on perennial sea-ice cover in the Arctic Ocean. Although the speleothems' record of this study also indicates intervals of permafrost before the formation of perennial sea ice at about 400 ka, such permafrost was prone to thawing in times of higher Northern Atlantic and/or global mean temperature. The long-term cooling of the Arctic Ocean that occurred between around 1,350 ka and around 400 ka eventually reached a temperature threshold for the formation of perennial Arctic sea ice, which stabilized the presence of continuous permafrost in Siberian regions where it remains today. This record indicates that, under future open-water Arctic scenarios, as predicted for later this century[30], this stabilization is likely to be removed, accelerating the northerly retreat of continuous permafrost.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1880-1.

1. IPCC *Climate Change 2013: The Physical Science Basis* (eds Stocker, T. F. et al.) (Cambridge Univ. Press, 2013).
2. Bell, D. B., Jung, S. J. A. & Kroon, D. The Plio-Pleistocene development of Atlantic deep-water circulation and its influence on climate trends. *Quat. Sci. Rev.* **123**, 265–282 (2015).
3. Cronin, T. M. et al. Enhanced Arctic amplification began at the Mid-Brunhes Event ~400,000 years ago. *Sci. Rep.* **7**, 14475 (2017).
4. Ballantyne, A. P. et al. The amplification of Arctic terrestrial surface temperatures by reduced sea-ice extent during the Pliocene. *Palaeogeogr. Palaeoclimatol. Palaeoecol.* **386**, 59–67 (2013).
5. Dai, A., Luo, D., Song, M. & Liu, J. Arctic amplification is caused by sea-ice loss under increasing CO_2. *Nat. Commun.* **10**, 121 (2019).
6. Screen, J. A. & Simmonds, I. The central role of diminishing sea ice in recent Arctic temperature amplification. *Nature* **464**, 1334–1337 (2010).
7. Lawrence, D. M., Slater, A. G., Tomas, R. A., Holland, M. M. & Deser, C. Accelerated Arctic land warming and permafrost degradation during rapid sea ice loss. *Geophys. Res. Lett.* **35**, L11506 (2008).
8. Liu, J., Curry, J. A., Wang, H., Song, M. & Horton, R. M. Impact of declining Arctic sea ice on winter snowfall. *Proc. Natl Acad. Sci. USA* **109**, 4074–4079 (2012).
9. Park, H., Walsh, J. E., Kim, Y., Nakai, T. & Ohata, T. The role of declining Arctic sea ice in recent decreasing terrestrial Arctic snow depths. *Polar Sci.* **7**, 174–187 (2013).
10. Zhang, T. Influence of the seasonal snow cover on the ground thermal regime: an overview. *Rev. Geophys.* **43**, RG4002 (2005).
11. Serreze, M. C. & Stroeve, J. Arctic sea ice trends, variability and implications for seasonal ice forecasting. *Phil. Trans. R. Soc. Lond. A* **373**, 20140159 (2015).
12. Schuur, E. A. G. et al. Vulnerability of permafrost carbon to climate change: implications for the global carbon cycle. *Bioscience* **58**, 701–714 (2008).
13. Anisimov, O. & Reneva, S. Permafrost and changing climate: the Russian perspective. *Ambio* **35**, 169–175 (2006).
14. Vaks, A. et al. Speleothems reveal 500,000-year history of Siberian permafrost. *Science* **340**, 183–186 (2013).
15. Lauriol, B., Ford, D. C., Cinq-Mars, J. & Morris, W. A. The chronology of speleothem deposition in northern Yukon and its relationships to permafrost. *Can. J. Earth Sci.* **34**, 902–911 (1997).
16. Brown, J., Ferrians, J. O. J., Heginbottom, J. A. & Melnikov, E. S. *Circum-Arctic Map of Permafrost and Ground-ice Conditions* Circum-Pacific Map Series CP-45 (US Geological Survey & Circum-Pacific Council for Energy and Mineral Resources, 1997).
17. Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen–Geiger climate classification. *Hydrol. Earth Syst. Sci.* **11**, 1633–1644 (2007).
18. Cheng, H. et al. Improvements in ^{230}Th dating, ^{230}Th and ^{234}U half-life values, and U–Th isotopic measurements by multi-collector inductively coupled plasma mass spectrometry. *Earth Planet. Sci. Lett.* **371–372**, 82–91 (2013).
19. Lisiecki, L. E. & Raymo, M. E. A Pliocene-Pleistocene stack of 57 globally distributed benthic δ^{18}O records. *Paleoceanography* **20**, PA1003 (2005).
20. Hansen, J. et al. Global temperature change. *Proc. Natl Acad. Sci. USA* **103**, 14288–14293 (2006).
21. Laskar, J. et al. A long-term numerical solution for the insolation quantities of the Earth. *Astron. Astrophys.* **428**, 261–285 (2004).
22. Kender, S. et al. Closure of the Bering Strait caused Mid-Pleistocene Transition cooling. *Nat. Commun.* **9**, 5386 (2018).
23. Maroon, E. A., Kay, J. E. & Karnauskas, K. B. Influence of the Atlantic meridional overturning circulation on the Northern Hemisphere surface temperature response to radiative forcing. *J. Clim.* **31**, 9207–9224 (2018).
24. Lawrence, K. T., Sosdian, S., White, H. E. & Rosenthal, Y. North Atlantic climate evolution through the Plio-Pleistocene climate transitions. *Earth Planet. Sci. Lett.* **300**, 329–342 (2010).
25. Cronin, T. M. et al. Quaternary ostracode and foraminiferal biostratigraphy and paleoceanography in the western Arctic Ocean. *Mar. Micropaleontol.* **111**, 118–133 (2014).
26. Vandenberghe, J. et al. Eurasian permafrost instability constrained by reduced sea-ice cover. *Quat. Sci. Rev.* **34**, 16–23 (2012).
27. Kurita, N. Origin of Arctic water vapor during the ice-growth season. *Geophys. Res. Lett.* **38**, (2011).
28. Bintanja, R. & Selten, F. M. Future increases in Arctic precipitation linked to local evaporation and sea-ice retreat. *Nature* **509**, 479–482 (2014).
29. O'Neill, H. B. & Burn, C. R. Impacts of variations in snow cover on permafrost stability, including simulated snow management, Dempster Highway, Peel Plateau, Northwest Territories. *Arctic Sci.* **3**, 150–178 (2017).
30. Barnhart, K. R., Miller, C. R., Overeem, I. & Kay, J. E. Mapping the future expansion of Arctic open water. *Nat. Clim. Change* **6**, 280–285 (2016).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2020

Methods

Description of the caves

Ledyanaya Lenskaya Cave. The cave is located 116 km east-southeast of the town of Lensk, 180 m above sea level, with the cave entrance located on the northeastern riverbank in a cliff about 50 m above the Lena River. The local vegetation is subboreal taiga forest[31]. The cave is developed in Ediacaran limestones and marls and its length is mapped to about 21 m. An approximately 90-m-long main passage ascends by about 15° in a north-northeast direction, ending in a central roughly 55-m-long and 10–20-m-wide hall, with a ceiling height up to 8 m. The hall is mostly filled with massive ice several metres thick. A narrow passage leads from its top to the cave's upper section. The latter is about 70 m long, ascending by about 20°, and consists of two chambers connected by a narrow passage and partly filled by ice. The walls of these chambers are partly covered by flowstone and stalactites (SLL14, taken in 2014; Extended Data Fig. 1A). Around 50 m from the main entrance, a narrow slightly ascending roughly 40-m-long corridor splits from the main passage in a east-northeast direction, and ends in small chambers in which vadose speleothems were collected (SLL9 and SLL10, taken in 2009–2010; Extended Data Fig. 1A). The depth of this chamber below the surface is 15–20 m, and the depth of the large hall is about 60 m.

The cave air temperature was monitored using HOBO UA Pendant Temperature Loggers from March 2010 to May 2013 (logger 'Siberia 7') and November 2013 (logger 'Siberia 6') (Extended Data Fig. 2A, B). The loggers' measurement uncertainty is ±0.47 °C. The temperature in the central hall with massive ice was measured with logger 'Siberia 6' (Extended Data Fig. 1A) and was found to vary between −0.1 °C and 0.0 °C in early spring, and between +0.6 °C and +0.7 °C in summer (Extended Data Fig. 2A). Logger 'Siberia 7' measured the temperature in the chamber with speleothems SLL9 and SLL10, and the temperature there is relatively stable at about +0.3 °C (Extended Data Fig. 2A). Although in both places temperatures are slightly above zero, no water seepage or speleothem growth was found beyond several metres from the entrance, showing that the rock above the cave is frozen year-round. In the uppermost cave chambers where SLL14 speleothem samples were collected, the temperature in January 2014 was 0 °C and the ice was dry. According to Lensk meteorological station (Extended Data Fig. 2B) the MAAT between 2010 and 2013 was about −6 °C, thus the cave is 6 °C warmer. The cave's ascending morphology, with the entrance being its topographically lowest point, causes warm humid air to be trapped inside during the summer months. Formation of ice in the uppermost parts of the cave shows that the cooling that creates the ice occurs when warm and moist air comes in contact with the subzero temperature of the cave walls[32]. This is also supported by the permafrost map, which shows continuous permafrost (type 18) above the Ledyanaya Lenskaya Cave[33] (Extended Data Fig. 3A). More information about the cave can be found in the Supplementary Online Materials of Vaks et al.[14].

Botovskaya Cave. The cave system is located 58 km north-northeast of the town of Zhigalovo, 750 m above sea level, at the head of a small valley northeast of the Boti River, which joins the Lena River 8.6 km southeast of the cave. The vegetation is subboreal taiga forest[31]. The cave is located in Ordovician limestones and sandstones, and is the longest cave system in Russia, reaching a total length of >69 km and a depth of up to 130 m below the surface, comprising a horizontal maze of thousands of passages developed along the criss-cross system of tectonic fissures (Extended Data Fig. 1B).

Cave air temperatures were monitored using HOBO UA Pendant Temp loggers, from February 2010 to February 2016. The temperatures in the deeper parts of the cave are stable at +1.9 ± 0.3 °C but vary from −0.2 °C to +1.6 °C near the entrance, with minima in winter and maxima in summer (Extended Data Fig. 2C). Surface air temperatures (SAT) outside the cave were also monitored by a HOBO logger tied on the northern shady side of a big tree 2 m from the ground. The mean annual SAT is −2.1 °C, varying between summer and winter extremes from +37 °C to <−40 °C (−40 °C is the minimum limit of the logger; during two nights in winter 2010–2011 the minimum temperature dropped below −40 °C). The MAAT is thus lower than the cave air temperature by 3–4 °C (Extended Data Fig. 2D).

Regional permafrost is discontinuous and found below the Boti River Valley and its slopes, but is absent from the plateau above the inner parts of Botovskaya Cave[33] (Extended Data Fig. 3B). The eastern part of the cave ('New World') is where the most water seepage and modern speleothem deposition occurs today (Extended Data Fig. 3B). This cave section is located below a small surface depression that hosts an intermittent stream that allows rain/snowmelt water to seep into the cave. Thermal energy from infiltrating water probably contributes to an absence of permafrost in this area. Speleothem samples were collected in the 'New World' section of the cave. The western part of the cave ('Old World') is drier, with some passages near the entrance clogged with massive ice.

Speleothem petrography

In Ledyanaya Lenskaya Cave, the speleothem cover on the cave walls and floor is usually 5–10 cm thick. Speleothems consist of several calcite horizons, each composed of brown or grey columnar calcite crystals usually free from detrital material (Extended Data Figs. 4, 6A, B). These calcite horizons are separated by whitish or beige thin (<2 mm) layers of microcrystalline calcite, sometimes containing pieces of marl and limestone host rock. These layers represent growth hiatuses, sometimes with pieces of broken host rock from the cave ceiling remaining on the ancient speleothem surface.

In Botovskaya Cave, speleothem deposition is much more widespread and massive than in Ledyanaya Lenskaya Cave. Here, the thickness of the speleothems is many tens of centimetres, showing that compared with Ledyanaya Lenskaya Cave, the humid and warmer climate of the area provided speleothems with better opportunities to grow (Extended Data Fig. 5). Active speleothems are also found in this cave. Unlike in Ledyanaya Lenskaya Cave, most speleothems in Botovskaya Cave are composed of aragonite (Extended Data Fig. 6C, D), but with some calcite speleothems (for example, stalagmite SB-6919) and some speleothems comprising alternate aragonite and calcite layers (for example, stalagmite SB-01112). Apparent growth breaks, sometimes separating calcite and aragonite layers, are common.

Methods used in this study

The speleothems were sectioned using a diamond saw. For the purpose of U–Th dating, between 10 and 250 mg of powder was drilled from each sampled horizon using 0.8–1.0 mm drill bits. Speleothem mineralogy was examined at ETH Zurich, Switzerland, using a Bruker, AXS D8 Advance powder XRD diffractometer, equipped with a scintillation counter and an automatic sampler. Macro and microscope inspection shows that all horizons chosen for dating had a typical columnar petrography in calcite and fibrous petrography in aragonite (Extended Data Fig. 6), with almost no voids or re-crystallization marks, suggesting that they probably maintained closed-system conditions for U-series chronology (except for some growth hiatuses, such as that in SB-01112; Extended Data Fig. 5).

U–Pb chronology

Analytical methodology. Ages were determined by isotope dilution using a mixed ^{236}U–^{204}Pb–^{230}Th spike[34] and a first-generation Nu Plasma multi-collector inductively coupled plasma mass spectrometer. Subsamples were cut using a small diamond saw and transferred to acid-cleaned (1–2 M HNO_3 for >3 d) 15 ml polypropylene bottles. The subsamples were then sonicated repeatedly in 18 MΩ cm water until no suspended fines were visible, rinsing between each wash. Subsamples were then acid cleaned twice in distilled 2% HNO_3 with sonication to remove any residual dirt. Following each wash, samples were thoroughly

washed with 18 MΩ cm water and sonicated to remove any residual acid and dislodged surface material. Each acid wash was removed before the acid reaction completed, to prevent adsorption of dissolved ions back on to the surface of the sample.

One to two drops (about 30 µl per drop) of spike were added directly to the acid-cleaned carbonate and gently agitated to mix as the spike fully dissolved the sample. Once the visible reaction was complete, the solution was diluted to about 15 ml with 18 MΩ cm water, thoroughly shaken to homogenize and then immediately analysed, with no pre-concentration of U and Pb.

Analyses followed a six-step routine. In steps 0, 1, 2 and 3, ^{208}Pb, ^{207}Pb, ^{206}Pb, ^{204}Pb + ^{204}Hg and ^{202}Hg were measured using three ion counters at 2 AMU spacing. The relative gains of the ion counters were determined by stepping ^{204}Pb + ^{204}Hg alternately into each collector. ^{202}Hg was monitored to correct for ^{204}Hg on ^{204}Pb. In steps 4 and 5, ^{238}U was measured on a Faraday cup, with ^{235}U and ^{236}U measured alternately on both a Faraday cup and an ion counter; this allows the use of the Faraday/Faraday ^{238}U/^{235}U ratio, or the Faraday/ion counter ^{238}U/^{235}U ratio depending on the ^{235}U signal intensity. Faraday/Faraday ratios were normally used for both ^{238}U/^{235}U and ^{238}U/^{236}U.

Mass fractionation was estimated using the measured ^{238}U/^{235}U ratio of the samples and an assumed natural value of 137.75 (ref. [35]). On the basis of previous testing of the instrument, the mass fractionation for Pb was assumed to be 2‰ per AMU higher than for U (ref. [36]).

Before the first analysis, the Nu Instruments DSN100 sample introduction system and sample lines were cleaned with 10% HNO_3, 2% HNO_3 and 18 MΩ cm water. A dedicated set of B-type cones reserved for very-low-level Pb work were used. These were gently cleaned by rinsing with deionized water before use to remove excessive Ca build-up from the skimmer orifice. As far as possible, the surface coating on the cones was not disturbed. The instrument was then initially tuned and optimized with a 100 ppt thallium (Tl) solution and diluted natural U solution. Intentional addition of Pb was avoided during tuning to prevent re-contamination of the instrument, but sufficient Pb blank is present in the Tl solution to see the Pb peaks on the ion counters. Peak shape and optimization was then re-checked on samples; focusing settings for the zoom optics often changed substantially from the nominally clean Tl solution to the matrix-heavy samples, especially following cleaning of the cones. Gas flows were also re-optimized to suppress interferences (probably from Sr_2O_2), which manifest as superimposed peaks about 0.2 AMU lighter than the Pb peaks, especially on ^{208}Pb. Optimization was checked again after an initial couple of sacrificial analyses and regularly during the analytical session. The DSN100 was re-cleaned with 18 MΩ cm water every 1–2 d to remove Ca build-up and U and Pb blank.

Separate sample aliquots up to about 0.2 g were dissolved and purified to obtain U cuts for measurement of the ^{234}U/^{238}U ratio. Purification used 2 ml columns with AG1X8 anion exchange resin. Samples were loaded in, and Ca eluted with about 10 M HCl. U was eluted with 18 MΩ cm water. The purified U was measured on the same instrument, with ^{234}U and ^{238}U measured on ion counter and Faraday collectors, respectively. Standard bracketing with CRM145 (CRM112A) was used to correct both for mass fractionation and ion-counter gain.

Non-radiogenic Pb correction. Model ages were calculated from each pair of U–Pb and ^{234}U/^{238}U analyses. The ^{238}U–^{206}Pb decay provides the age data used here, but the ^{235}U–^{207}Pb system was also measured and provides an assessment of concordance, and thus confidence in obtained ages.

Ages were calculated using an estimated ^{208}Pb/^{206}Pb (and ^{208}Pb/^{207}Pb) ratio for the initial non-radiogenic Pb and, the modern-day measured disequilibrium in the ^{234}U/^{238}U ratio to constrain the initial ^{234}U/^{238}U ratio. ^{208}Pb is assumed to be entirely non-radiogenic on the basis that ^{232}Th is typically at very low concentration in speleothems and that samples are young compared with the ^{232}Th half-life. Common ^{208}Pb/^{206}Pb (and ^{208}Pb/^{207}Pb) for the non-radiogenic Pb correction was estimated by a combination of:

(1) Identifying and analysing unradiogenic parts of the sample, not greatly modified by Pb from decay.

(2) Retrospectively picking approximate isochrons from the data, on the basis that given a large number of analyses the following are likely: (a) age overlap between samples/subsamples can be expected and hence some clumping of analyses along mixing lines between the initial Pb composition and the radiogenic composition for a given (approximate) age; (b) data should fan around the initial Pb composition; (c) subsamples of roughly the same age, can, to a first approximation be grouped based on the observed ^{234}U/^{238}U ratio. The latter is somewhat limited by initial ^{234}U/^{238}U variation, but surviving ^{234}U disequilibrium decreases by a factor of 2 for each ^{234}U half-life, so for a spread of ages over a few hundred thousand years or more, the variations due to decay of ^{234}Uxs will dominate over variations in the initial ratio.

(3) Linear regressions[35,36] through groups of data of similar age in ^{234}U(or ^{238}U)/^{206}Pb – ^{208}Pb/^{206}Pb isotope space allow an estimate of the initial ^{208}Pb/^{206}Pb ratio (^{235}U/^{207}Pb – ^{208}Pb/^{207}Pb isotope space for the initial ^{208}Pb/^{207}Pb ratio). Only groups containing relatively non-radiogenic analyses (ideally stratigraphically bracketed by more radiogenic analyses) were used, to minimize the effect of incorrectly grouping samples of different age. The groupings used and regression results are shown in Supplementary Table 1 and illustrated in Extended Data Fig. 7.

A common ^{208}Pb/^{206}Pb ratio of 1.471 ± 0.100 (and ^{208}Pb/^{207}Pb ratio of 2.465 ± 0.136) (95% confidence) was used for the Ledyanaya Lenskaya Cave samples (Extended Data Fig. 7). The former value is based mainly on a single grouping of samples that include the least-radiogenic Ledyanaya Lenskaya analysis, but is in agreement with a second generally more-radiogenic grouping. All other Ledyanaya Lenskaya data (except data rejected in Supplementary Table 1) are consistent with this common ^{208}Pb/^{206}Pb ratio (Extended Data Fig. 7).

For Botovskaya Cave samples, which include highly non-radiogenic material, a common ^{208}Pb/^{206}Pb ratio of 1.997 ± 0.213 (and ^{208}Pb/^{207}Pb ratio of 2.419 ± 0.123) (95% confidence) was determined in a similar way. As the Botovskaya data also includes some analyses that are almost entirely non-radiogenic, these have also been taken into account. Sample groupings and regression results are again shown in Extended Data Fig. 7 and Supplementary Table 1.

The rather different values of the common Pb composition between the two caves are probably attributable to host-rock composition. Botovskaya and Ledyanaya Lenskaya caves are hosted in rocks of Ordovician and Late Proterozoic age, respectively, which have had long periods to evolve distinctive Pb compositions before the formation of the speleothems they now host.

The use of model ages involves some degree of assumption about the uniformity of the common Pb composition. Consequently, an indication of the sensitivity of a particular age to the common Pb correction is given in Supplementary Table 1 and Extended Data Fig. 7, and found to be small compared with quoted uncertainties.

Pb blanks have not been separately corrected for and are dealt with as part of the total correction for non-radiogenic Pb. Given that a number of analyses yielded >99% radiogenic ^{206}Pb, the Pb blank can be considered a generally minor source of non-radiogenic Pb. Sample ^{204}Pb was corrected in the isotope dilution calculation using ^{208}Pb as a proxy, assuming a natural ^{208}Pb/^{204}Pb ratio of 37.1 ± 10 (95% confidence). For most analyses, >99% of the total ^{204}Pb originated from the tracer, so the correction is small.

Ages were calculated from the common-Pb-corrected ^{238}U/^{206}Pb (and ^{235}U/^{207}Pb) ratio and measured ^{234}U/^{238}U ratio using in-house software. Uncertainties were estimated using a Monte Carlo approach. Initial ^{230}Th and ^{231}Pa are assumed to have been absent in the age/concordia calculations. Initial ^{234}U is determined based on the present-day ^{234}U/^{238}U ratio as part of the age calculation, much as in the U–Th chronology.

The decay constants used are: ^{238}U, 1.55125×10^{-10}; ^{234}U, 2.82203×10^{-6}; ^{230}Th, 9.17055×10^{-6}; ^{226}Ra, 4.33488×10^{-4}; ^{235}U, 9.8485×10^{-10}; ^{231}Pa, 2.11583×10^{-5}; ^{232}Th, 4.9475×10^{-11} (refs. [18,37,38]).

As a cross-check on this methodology, two layers of stalagmite SLL10-6, G and F, were also analysed following the method of Mason et al.[34], which involved chemical purification of Pb from subsamples to generate isochron ages. Within uncertainty, these agree with the analyses obtained with the simplified protocol outlined above. U–Pb ages were also cross-checked against U–Th ages determined for the parts of the samples <0.5 Ma old (Supplementary Tables 1, 2, Extended Data Fig. 8).

Screening of data and data quality

Ledyanaya Lenskaya Cave. Of the 59 U–Pb analyses (Supplementary Table 1), ^{238}U–^{234}U–^{206}Pb ages were included from 52 analyses (Extended Data Fig. 8A). Of the excluded analyses, one has no corresponding $^{234}U/^{238}U$ measurement (SLL10-4-B bottom), so an age cannot be calculated. Three analyses (two from SLL10-6-B and one from SLL10-6-B/C hiatus) are non-radiogenic and yield ages with unhelpfully large uncertainties, though usefully help to constrain the common Pb correction. Two analyses (SLL14-1-C centre and SLL9-1-A2) yield >1.5 Ma apparent ages, which are not reproduced in other analyses and appear inconsistent with the $^{234}U/^{238}U$ measurements; and one analysis (SLL10-4-A top) is both out of stratigraphic order and in disagreement with two other analyses from the same layer.

The remaining 52 ages are consistently in stratigraphic order. The about 1.3–1.5 Ma age cluster is useful for demonstrating analytical robustness, since the common Pb correction in these analyses is mostly small and layers can be dated in stratigraphic order with about 40 ka age resolution, even without pre-concentration of Pb. The replication of ages of many growth periods between different samples, even where the magnitude of the common Pb correction varies substantially, helps to validate the common Pb value used to calculate the model ages and corroborates the analytical robustness of the method. The agreement of the ages for SLL10-6-F top and SLL10-6-G with and without pre-concentration of U and Pb should also be noted (Extended Data Fig. 8A).

^{235}U–^{207}Pb ages are also given for reference in Supplementary Table 1 and are mostly concordant with the ^{238}U–^{234}U–^{206}Pb ages within error (Extended Data Fig. 9) (or, in five cases, show only slight discordance). These small discordances can probably be attributed to uncertainties in the common Pb correction, and could be accounted for by shifts in the common Pb composition of a few percent.

Samples SLL10-4C, SLL10-5B, SLL10-9A, SLL10-9C and SLL14-1B bottom have notably low ^{235}U–^{207}Pb ages relative to their ^{238}U–^{234}U–^{206}Pb ages. None of these ^{238}U–^{234}U–^{206}Pb ages fails on the grounds of being inconsistent with stratigraphic position and they tend to replicate well in other samples (Extended Data Fig. 8). It is likely that this discordance is an analytical artefact, specifically incomplete elimination of the molecular interference on ^{208}Pb. An extraneous contribution on ^{208}Pb will lead to overcorrection of the common Pb. Since ^{207}Pb is typically some 20–30 times more sensitive (Supplementary Table 1) to the common Pb correction, a small overcorrection on the ^{206}Pb can correspond to a significant overcorrection on ^{207}Pb, hence the ^{235}U–^{207}Pb ages appear anomalously young while the ^{238}U–^{234}U–^{206}Pb ages remain stratigraphically consistent.

The (near-)concordance of the majority of the ages provides confidence in the data from an analytical standpoint.

Botovskaya Cave <0.7 Ma samples. Of the 13 analyses from Botovskaya Cave, 12 ^{238}U–^{234}U–^{206}Pb ages were calculated. The analysis that failed to produce an age was from non-radiogenic calcite G layer in SB-01112, which is bracketed between more radiogenic aragonite layers, and so it still useful for constraining the common Pb composition. Ages fall between 0.4 and 0.7 Ma (Supplementary Table 1), and are in stratigraphic order overlapping well with available U–Th ages (Extended Data Fig. 8B). ^{235}U–^{207}Pb ages could only be calculated for four analyses and are of low precision, but they are all concordant with the ^{238}U–^{234}U–^{206}Pb ages (Supplementary Table 1).

Botovskaya Cave pre-0.7 Ma samples. Botovskaya Cave shows an extensive 2 Ma and older record, which is the subject of ongoing investigation. The data of some of these samples are used to better constrain the common Pb correction in the present work (Supplementary Table 1, Extended Data Fig. 7), although the ages are not used here because no pre-0.7 Ma Botovskaya samples overlapping the older part of the Ledyanaya Lenskaya record have yet been identified.

U–Th chronology. The U and Th analysis, as well as U–Th ages calculations were performed following the Supplementary Online Methods of Vaks et al.[14]. Twelve layers of five samples from Botovskaya Cave were dated. Dating results and their correction for initial Th were calculated using Isoplot 4.15[39] and are presented in Supplementary Table 2. ^{238}U concentrations in calcite vary between 0.44 and 1.25 ppm, whereas in aragonite they vary between 40.4 and 136.7 ppm. ^{232}Th concentrations usually vary between 0.06 and 3.15 ppb. $^{230}Th/^{232}Th$ activity ratios in the analysed samples varied between 6,322 and >10^6, and the $^{232}Th/^{238}U$ activity ratios varied between 4.78×10^{-4} and 8.94×10^{-7}. Thus, correction for $Th_{initial}$ was negligible. Two ages from layer SB-01112-D were rejected because this layer is adjacent to a horizon with high porosity, representing a joint between a growth hiatus and a crack in the speleothem. The lower part of layer 'SB-01112-D bottom' shows an age reversal, possibly due to U leaching. Other U–Th ages shown in Extended Data Figs. 4, 5, 8 are taken from Vaks et al.[14], slightly modified using updated half-lives of ^{234}U and ^{230}Th (ref. [18]).

The ability of Arctic air to reach Ledyanaya Lenskaya Cave site

Arctic air masses bring polar air southward to substantially influence Siberian weather. In particular, these air masses are capable of bringing latent heat and moisture from the unfrozen Arctic Ocean into the Siberian continental interior during the period of October–early November. During this time of the year, the Arctic Ocean is still partly unfrozen and snow cover is forming over the Siberian landmass. To demonstrate that these incursions of Arctic air are capable of reaching the site of Ledyanaya Lenskaya Cave, we show 12 examples of such snow events above Ledyanaya Lenskaya Cave in Extended Data Fig. 10. Three trajectories for air at elevations of 500, 1,500 and 5,000 m were calculated for each event with at least one trajectory in each event starting in the Arctic Ocean. These trajectories were assessed using the NOAA Hysplit program (https://ready.arl.noaa.gov/HYSPLIT.php)[40,41]. The proper timing of the weather events mentioned above was received from log data of the Lensk meteorological station (number 24923), 139 km northwest from Ledyanaya Lenskaya Cave (https://rp5.ru/Погода_в_Ленске,_Республика_Саха_(Якутия) (in Russian)[42,43]).

Data availability

The chronology data from this article are available at the NGDC database, at https://doi.org/10.5285/d56addfe-b7b8-4b70-9617-fadc1f7666b4. All data that support the findings of this study are available from the corresponding author on request as well.

31. Tchebakova, N. M., Rehfeldt, G. E. & Parfenova, E. I. in *Permafrost Ecosystems: Ecological Studies (Analysis and Synthesis)* Vol. 209 (eds Osawa, A. et al.) 427–446 (Springer, 2010).
32. Ford, D. & Williams, P. W. *Karst Geomorphology and Hydrology* (Unwin Hyman, 1989).
33. Seligman, B. The English language edition of the geocryological map of Russia and neighbouring republics. P.J. Williams and I.M.T. Warren. 1999. Ottawa: Collaborative Map Project. Handbook: 32 p, soft cover. ISBN 0-9685013-0-3. Map: 16 sheets (95 × 66 cm per sheet), scale 1:2,500,000. $US350.00. *Polar Record* **36**, 259–260 (2000).
34. Mason, A. J., Henderson, G. M. & Vaks, A. An acetic acid-based extraction protocol for the recovery of U, Th and Pb from calcium carbonates for U-(Th)-Pb geochronology. *Geostand. Geoanal. Res.* **37**, 261–275 (2013).
35. Hiess, J., Condon, D. J., McLean, N. & Noble, S. R. $^{238}U/^{235}U$ systematics in terrestrial uranium-bearing minerals. *Science* **335**, 1610–1614 (2012).
36. Mason, A. J. & Henderson, G. M. Correction of multi-collector-ICP-MS instrumental biases in high-precision uranium-thorium chronology. *Int. J. Mass Spectrom.* **295**, 26–35 (2010).

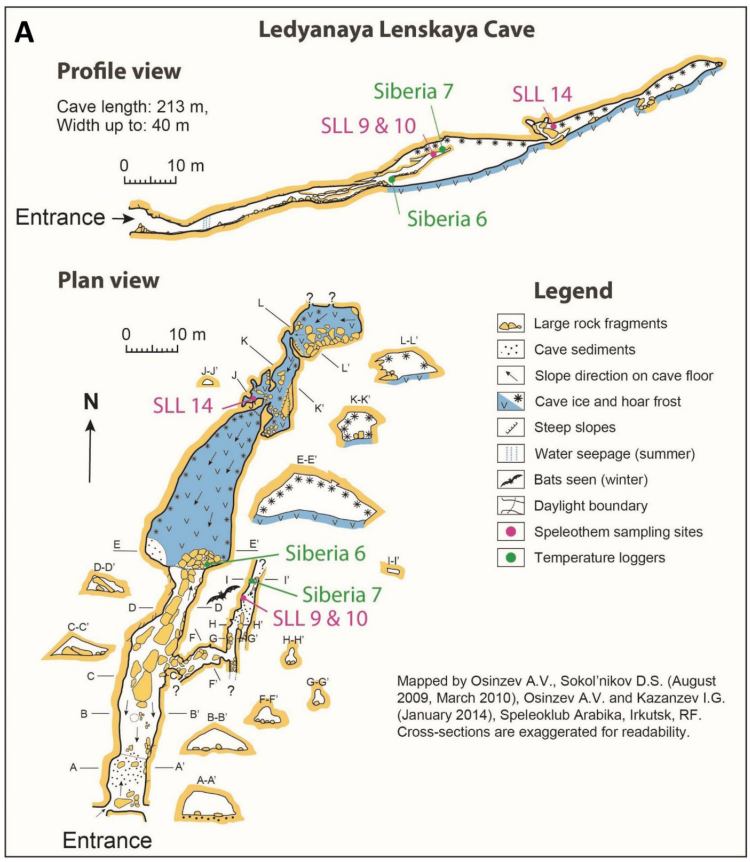
37. Cheng, H., Edwards, R. L., Murrell, M. T. & Benjamin, T. M. Uranium-thorium-protactinium dating systematics. *Geochim. Cosmochim. Acta* **62**, 3437–3452 (1998).
38. Steiger, R. H. & Jäger, E. Subcommission on geochronology: convention on the use of decay constants in geo- and cosmochronology. *Earth Planet. Sci. Lett.* **36**, 359–362 (1977).
39. Ludwig, K. A. (ed.) *Berkeley Geochronological Centre* Special Publication Vol. 5 (Berkeley Geochronological Centre, 2012).
40. Rolph, G., Stein, A. & Stunder, B. Real-time Environmental Applications and Display sYstem: READY. *Environ. Model. Softw.* **95**, 210–228 (2017).
41. Stein, A. F. et al. NOAA's HYSPLIT atmospheric transport and dispersion modeling system. *Bull. Am. Meteorol. Soc.* **96**, 2059–2077 (2015).
42. Bulygina, O. N., Veselov, V. M., Razuvaev, V. N. & Aleksandrova, T. M. *Baseline Climatological Data Sets* (RIHMI-WDC, 2014); http://meteo.ru/english/climate/cl_data.php.
43. Bulygina, O. N., Veselov, V. M., Razuvaev, V. N. & Aleksandrova, T. M. Description of hourly data set of meteorological parameters obtained at Russian stations. *Meteo* http://meteo.ru/english/climate/descrip12.htm (2014).

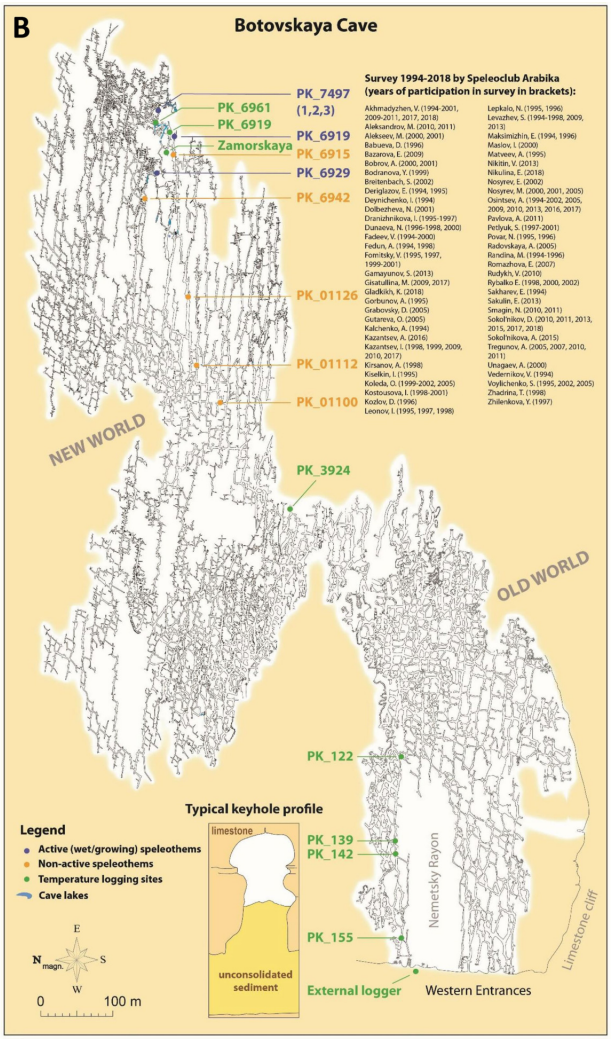
Acknowledgements We thank I. G. Kazanzev, V. A. Stol'arov, D. S. Sokol'nikov and other Speleoclub Arabica members for their help with expeditions' organizations, mapping the caves and collecting samples. We thank S. V. Alexeev, E. A. Kozireva, V. A. Pellinen and A. A. Svetlakov from Institute of Earth's Crust of the Russian Academy of Sciences (IEC RAS) in Irkutsk for help in the field and with samples shipment. We thank V. A. Balaev, and V. V. Alexioglo from Lensk, for their help during the fieldwork in Ledyanaya Lenskaya Cave. We thank O. Kwiecien from Northumbria University, Newcastle upon Tyne, F. Lechleitner from the University of Oxford, UK, and N. Brall from University of Lyon, France, for help with samples collecting in Botovskaya Cave. We thank B. Cohen from the Geological Survey of Israel for the help with graphic work. We thank M. Deininger, R. Parrish and L. Schirrmeister for useful comments that helped to improve the quality of this paper. We also acknowledge the NOAA Air Resources Laboratory (ARL) for the provision of the HYSPLIT transport and dispersion model and/or READY website (https://www.ready.noaa.gov) used in this publication. This work was funded by NERC Fellowship NE/G013829/1, NERC Standard Grant NE/K005057/1, Royal Society grant JP080831 and RFBR joint grant 09-05-92605 KO_a.

Author contributions A.V., G.M.H. and S.F.M.B. devised the approach of using caves to reconstruct permafrost, and raised the funding for this research. A.V., A.V.O., S.F.M.B. and O.S.G. conducted the field work, with help from A.M.K. A.J.M. led the development and application of the chronological work, with input from A.V. and G.M.H. A.V.O., M.R., A.B., S.F.M.B. and A.M.K. drew the maps (permafrost and caves) with input from all authors. A.V. led the interpretation of the data, and the writing of the manuscript, with input from all authors.

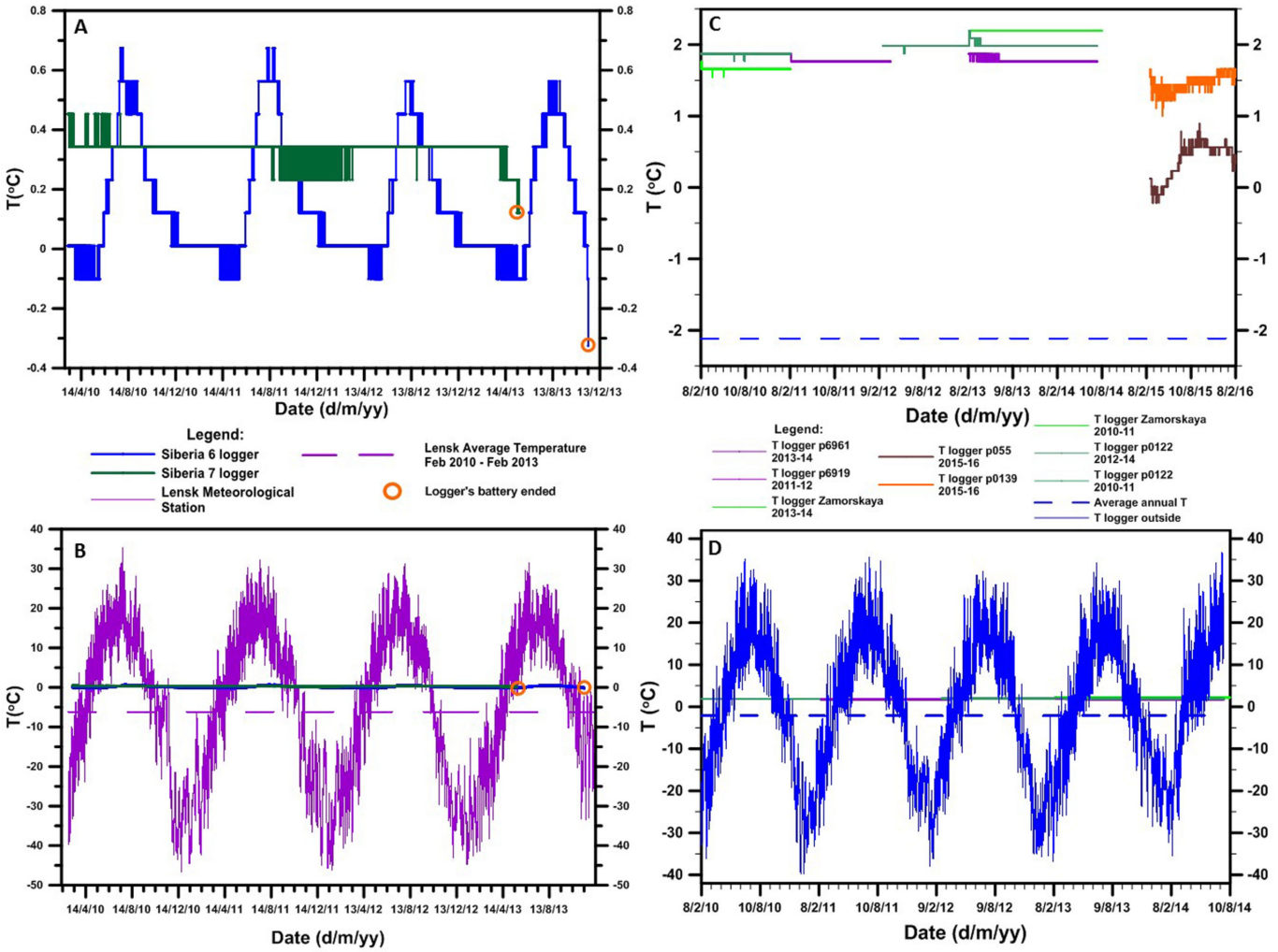
Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1880-1.
Correspondence and requests for materials should be addressed to A.V.
Peer review information *Nature* thanks Michael Deininger, Randall Parrish and Lutz Schirrmeister for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.

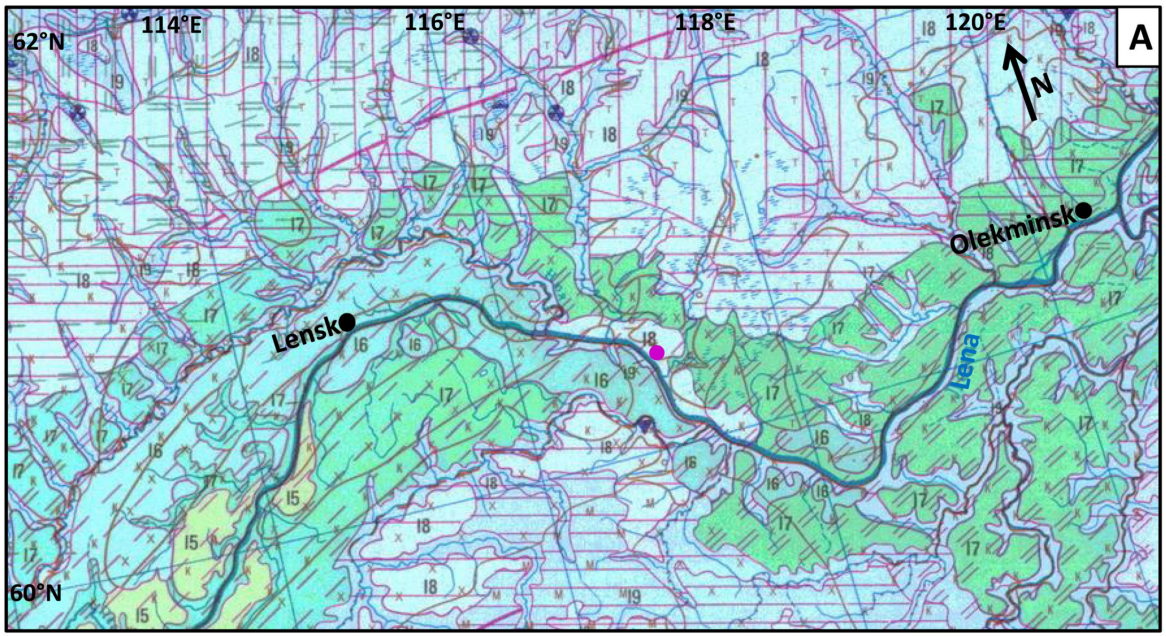


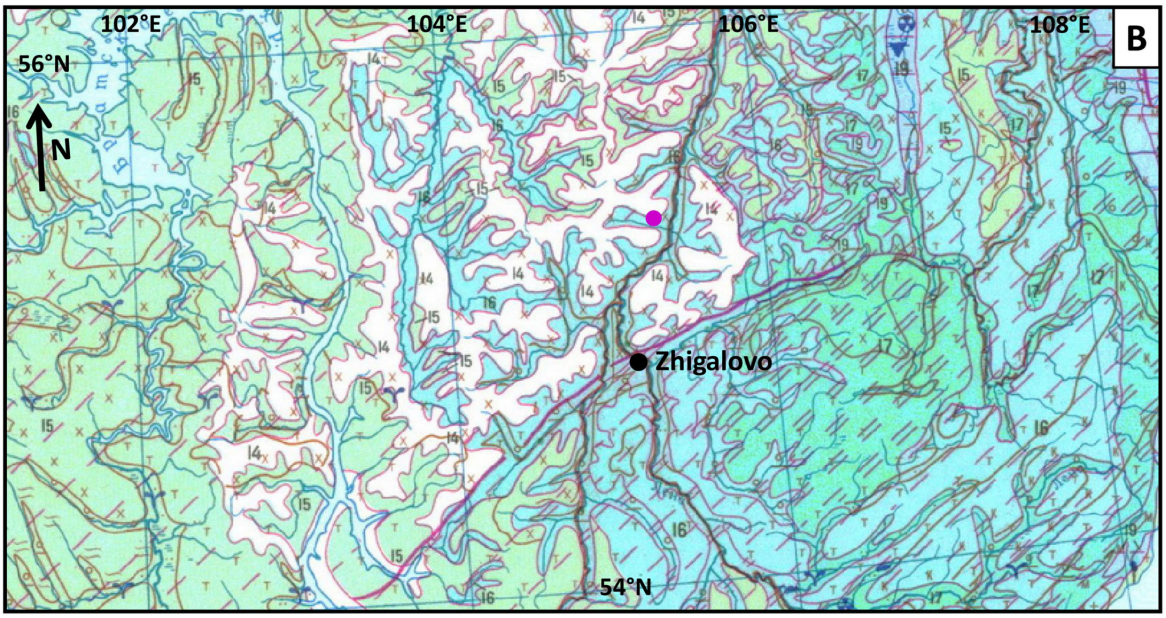


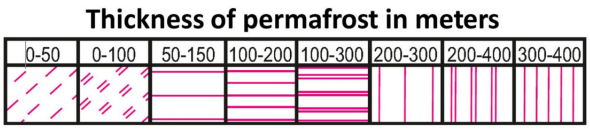
Extended Data Fig. 1 | Caves' maps. A, Map and cross-section of Ledyanaya Lenskaya Cave, with speleothem and temperature logger positions. **B**, Map of Botovskaya Cave with locations of speleothems and temperature loggers.

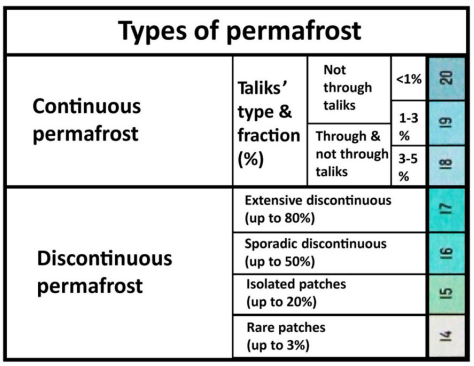


Extended Data Fig. 2 | Temperature data inside and outside caves. **A**, Ledyanaya Lenskaya Cave air temperatures between March 2010 and December 2013. **B**, Comparison of air temperature in Ledyanaya Lenskaya Cave (from **A**), mean annual surface temperature and surface air temperature changes (Lensk meteorological station 24923 data). **C**, Monitoring of air temperatures inside Botovskaya Cave (February 2010–February 2016). Between 2010 and 2014, temperatures were monitored deep inside the cave, whereas between 2015 and 2016 loggers were placed closer to the entrance. **D**, Comparison of Botovskaya Cave temperatures with mean surface temperature and surface temperature changes between February 2010 and July 2014 (data from temperature logger outside the cave).

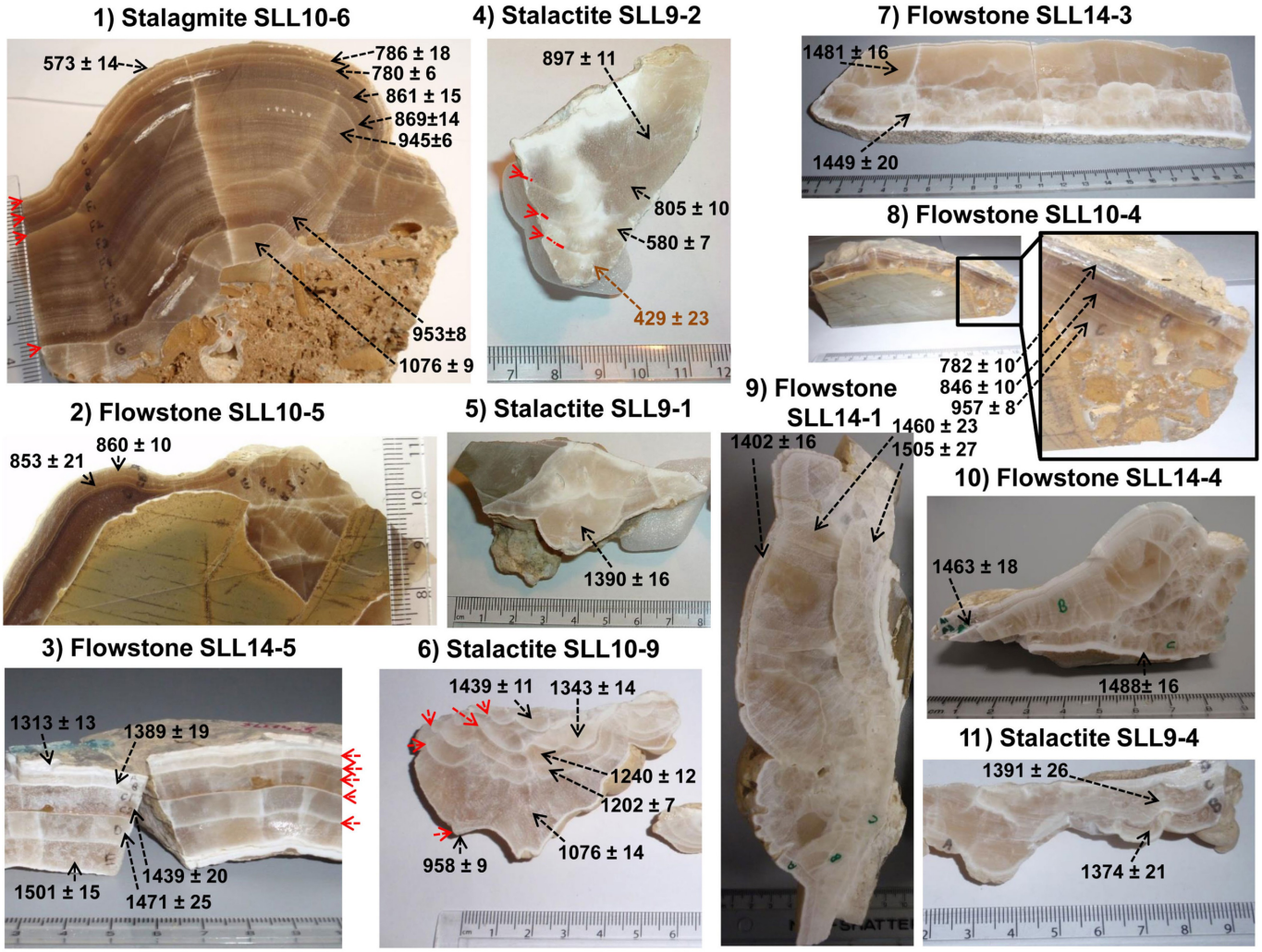




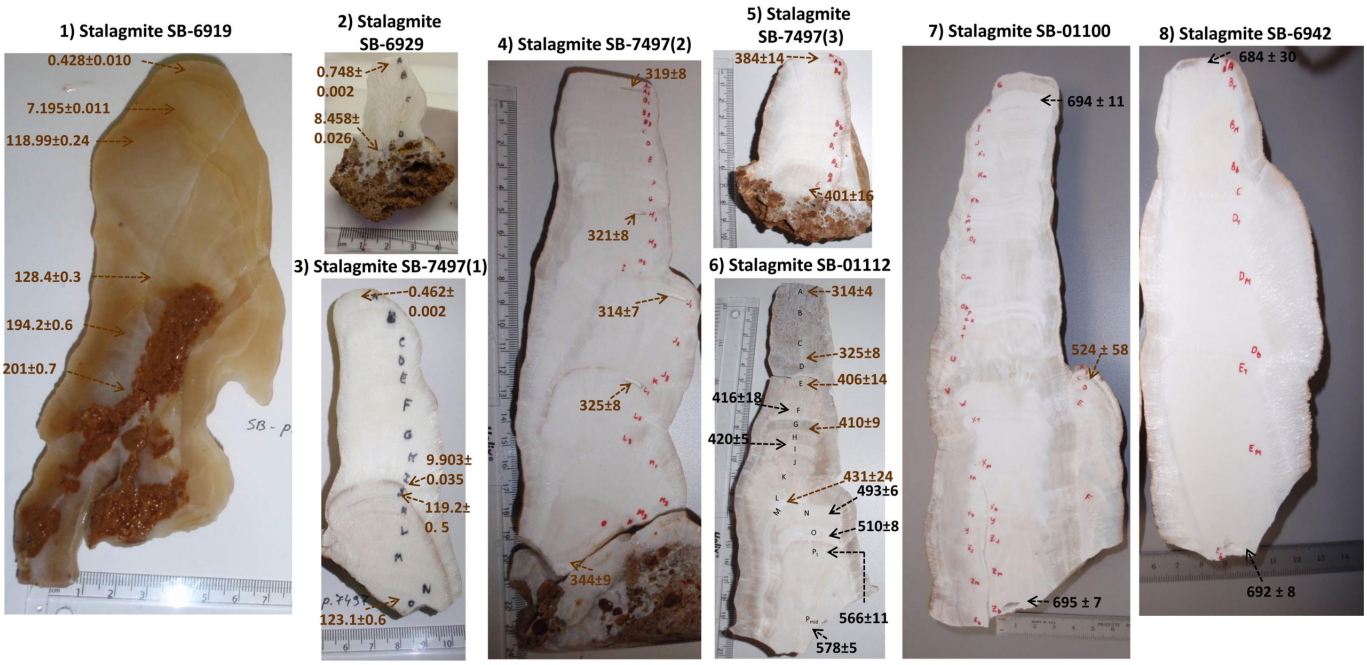




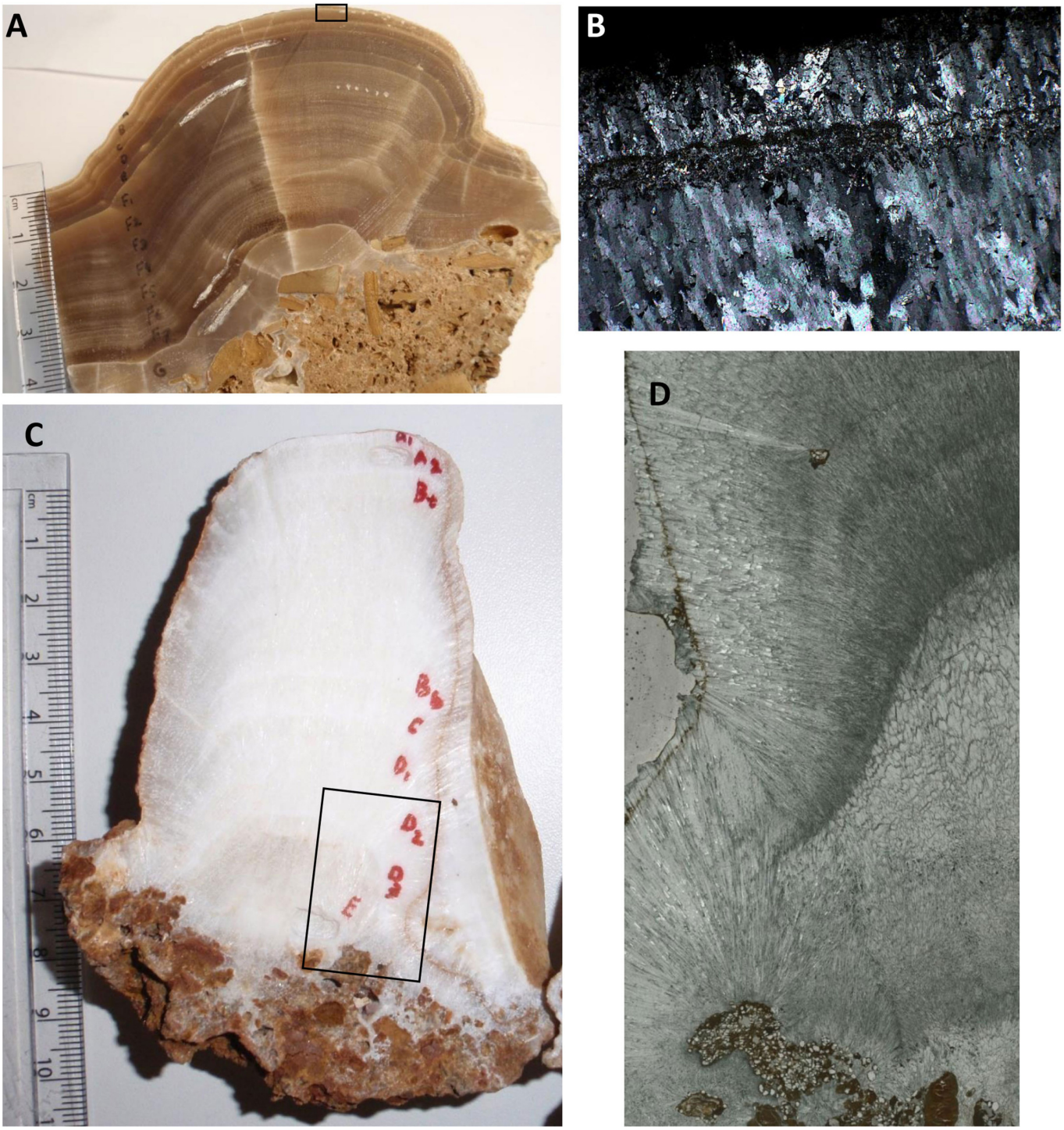
Extended Data Fig. 3 | Detailed permafrost maps. A, Permafrost map of the Ledyanaya Lenskaya Cave area (reproduced with permission from ref. [33], Cambridge Univ. Press). The cave site is marked by a magenta circle. Lensk and Olekminsk are marked by black circles. **B**, Permafrost map of the Botovskaya Cave area[33]. The cave location is marked by magenta circle. Zhigalovo is marked by black circle. Types and thickness of the permafrost are shown below. In places with continuous permafrost (types 18–20), it covers >95% of the area, but may contain small unfrozen units (taliks), mainly under permanent bodies of water. Taliks may go through the entire permafrost layer (through taliks), or through part of it (not through taliks).



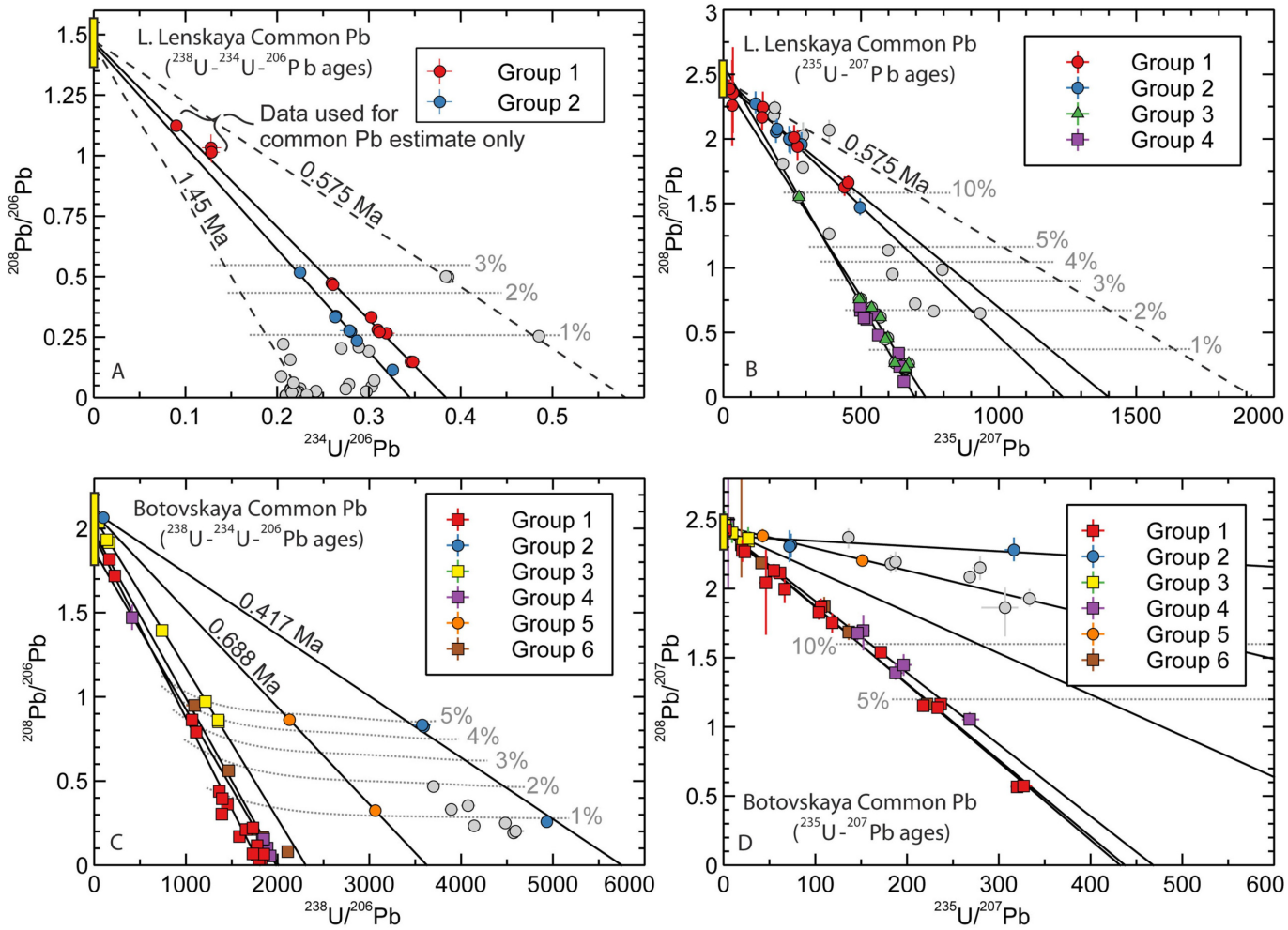
Extended Data Fig. 4 | Pictures of speleothems' cross-sections with ages (in ka) from Ledyanaya Lenskaya Cave. Age uncertainties are 2σ. The U–Pb ages shown in black and a single U–Th age in stalactite SLL9-2[14] is shown in brown. Hiatuses are shown with red arrows. All speleothems are composed of calcite.



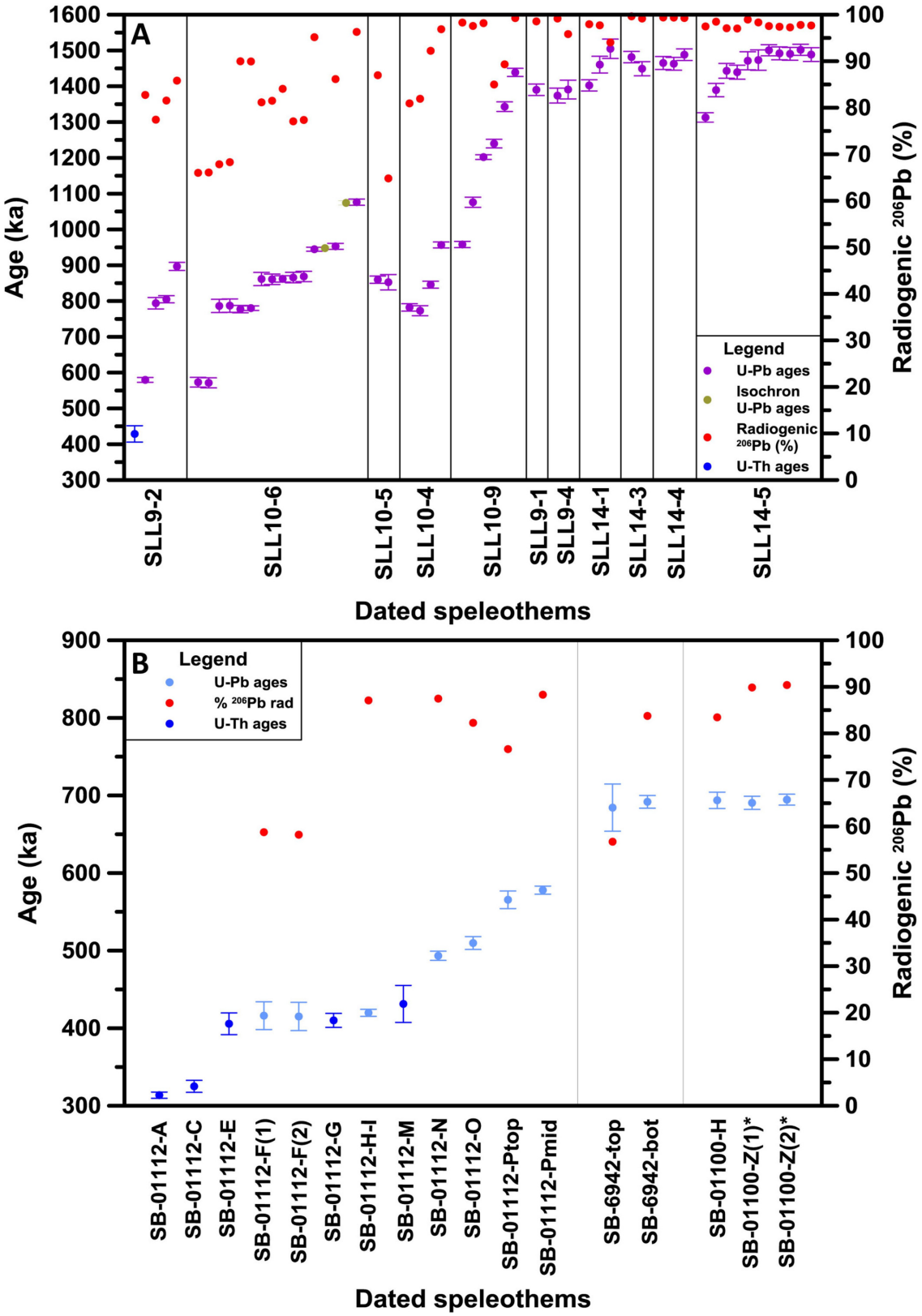
Extended Data Fig. 5 | Pictures of speleothems' cross-sections with ages (in ka) from Botovskaya Cave. Age uncertainties are 2σ. The U–Pb ages shown in black and U–Th ages[14] are shown in brown. Most speleothems are composed of aragonite, except stalagmite SB-6919, and layers A–D and G in stalagmite SB-01112, which are calcitic.



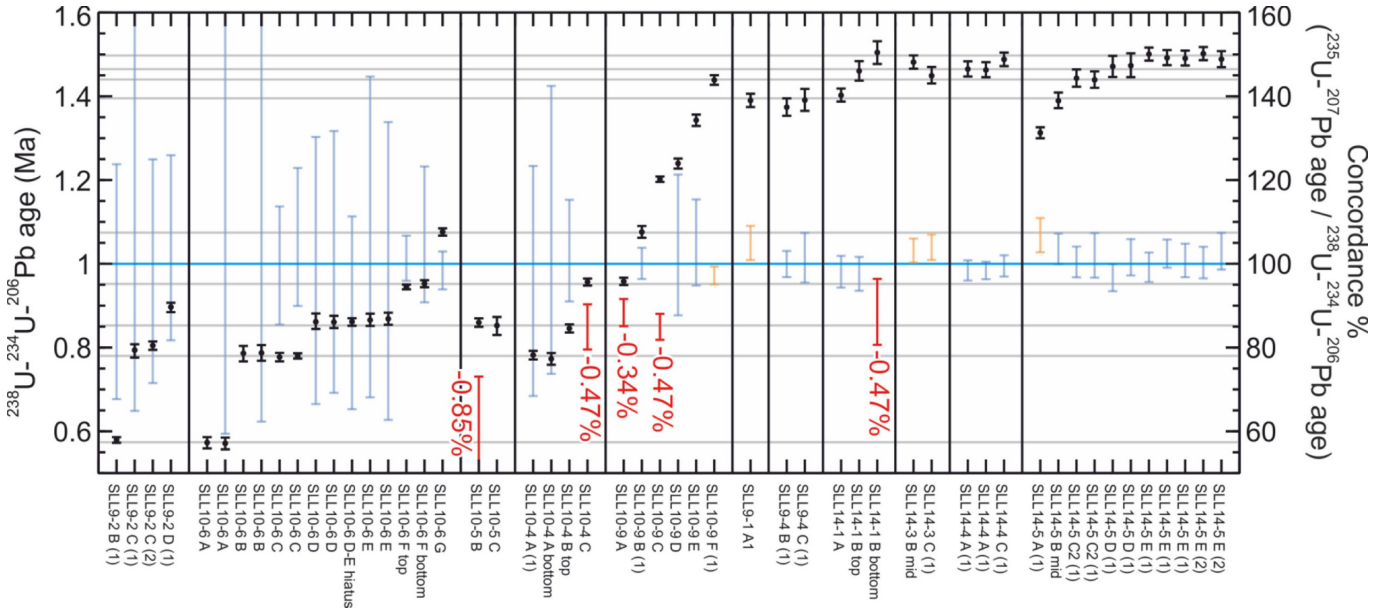
Extended Data Fig. 6 | Speleothems' petrography. A, Calcitic stalagmite SLL10-6 with its layered structure. **B**, Magnified area (5 to 3 mm—rectangle on the top of **A**) in crossed polar light, showing columnar crystals. **C**, Aragonitic stalagmite SB-7497(3) with fibrous crystals. **D**, Magnified area (3 to 1.5 cm) shown by black rectangle in **C** in plain polar light.



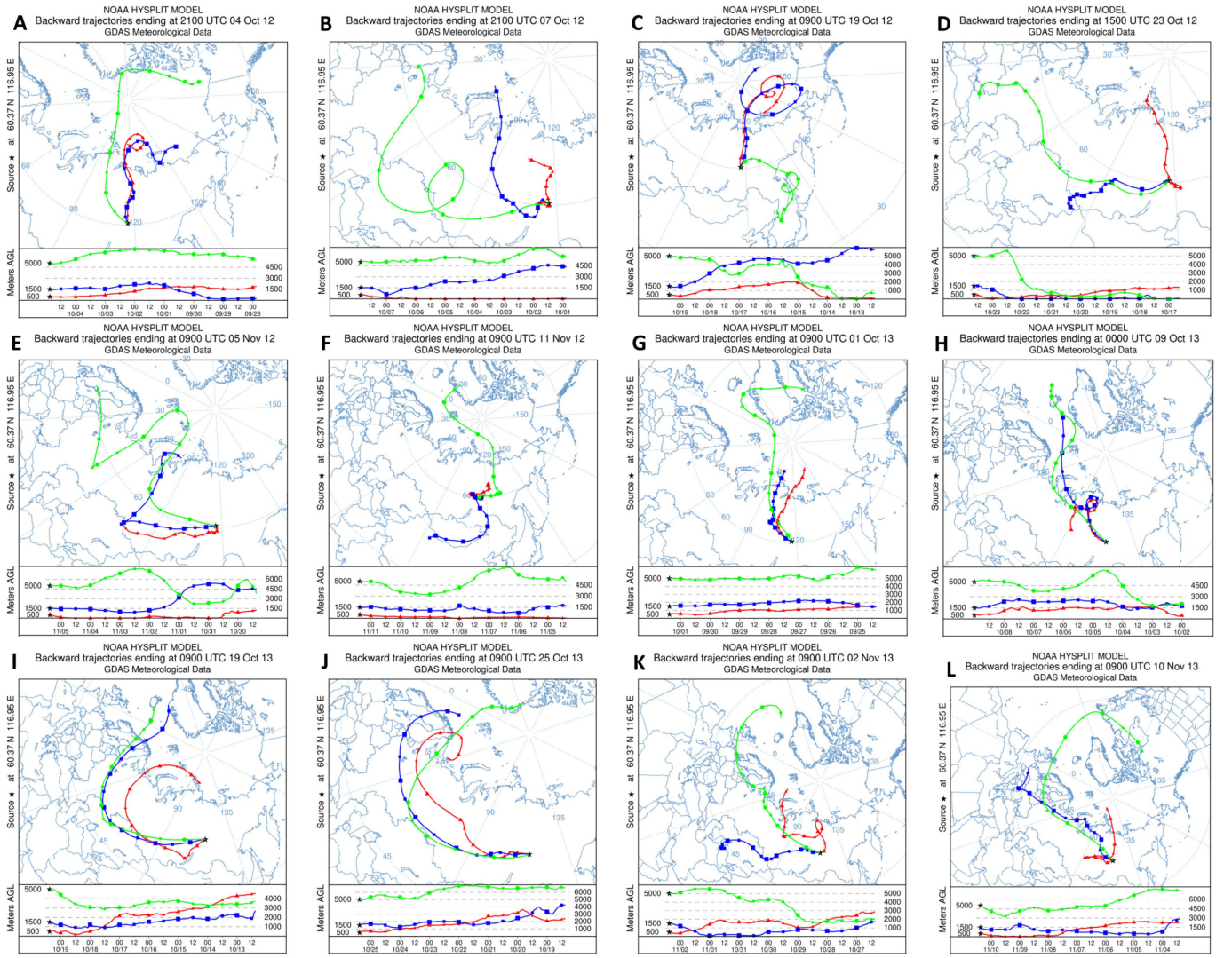
Extended Data Fig. 7 | Common Pb composition assessment for Ledyanaya Lenskaya (upper) and Botovskaya (lower) caves. ^{234}U–^{206}Pb and ^{238}U–^{206}Pb appear on the left, and ^{235}U–^{207}Pb on the right, showing consistent common Pb values for each cave. The *y* intercept represents the common lead ratio; yellow bars show the assigned range for common Pb in age calculations. Groups refer to the regression groups in Supplementary Table 1 with solid black lines showing regression fits. Grey contours show the percentage change in calculated age resulting from changing the common Pb composition by its assigned uncertainty. Dashed black lines show reference isochrons and ungrouped data are shown by grey circles. Uncertainties (2σ) are sometimes smaller than the symbol size. Particular details for each plot are as follows. **A**, Estimate of the common $^{208}Pb/^{206}Pb$ for Ledyanaya Lenskaya Cave in $^{234}U/^{206}Pb$–$^{208}Pb/^{206}Pb$ isotope space. ^{234}U is used in the plot instead of ^{238}U to suppress scatter in the $U/^{206}Pb$ ratio due to variations in the $^{234}U/^{238}U_{initial}$ ratio. All but two of the ungrouped data have age uncertainties due to common Pb of <1% and are thus insensitive to the common Pb correction. Most data that are more sensitive to the common Pb composition (that is, those between the 1% and 3% contours) are included in the regressions to estimate the common Pb composition. **B**, Equivalent plot in $^{235}U/^{207}Pb$–$^{208}Pb/^{207}Pb$ isotope space for the estimation of the common $^{208}Pb/^{207}Pb$ ratio for Ledyanaya Lenskaya cave. Groups 3 and 4 correspond to the clump of ungrouped data close to the horizontal axis at about 0.225 on **A**. Note that $^{235}U/^{207}Pb$ ages are substantially more sensitive to the common Pb correction; these ages are used as a check on U–Pb concordance rather than to derive the dates used in the paper. **C**, Estimate of the common $^{208}Pb/^{206}Pb$ for Botovskaya Cave in $^{238}U/^{206}Pb$–$^{208}Pb/^{206}Pb$ isotope space. Groups 1, 3, 4 and 6 (plotted as squares) are unpublished data from about 2 Ma samples included here only to provide additional constraint on the common Pb composition (Supplementary Table 1). Ages used here are from the data plotted as circles. As for Ledyanaya Lenskaya Cave, the common Pb estimate includes the data for which the correction matters most as far as is possible. **D**, Equivalent plot in $^{235}U/^{207}Pb$–$^{208}Pb/^{207}Pb$ isotope space for the estimation of the common $^{208}Pb/^{207}Pb$ ratio for Botovskaya Cave. $^{235}U/^{207}Pb$ data are more sensitive to the common Pb correction.



Extended Data Fig. 8 | Detailed speleothems' chronologies. A, The chronology of Ledyanaya Lenskaya Cave speleothems with 95% confidence errors. Each column in the plot represents one individual speleothem named in bottom horizontal axis. In each column, the U–Pb ages (purple circles) appear in stratigraphic order from left (young) to right (old). For each U–Pb age, the corresponding proportion of radiogenic ^{206}Pb (right vertical axis) is shown by red circles above. The U–Th age of the youngest layer A in the stalactite SLL9-2[14] is shown by blue circle (bottom left). The two isochron ages of layers F top and G in SLL10-6 stalagmite are shown by olive circles. Several replicate age determinations for similar layers were performed and appear in the plot in the same order as shown in Supplementary Table 1. **B**, Botovskaya Cave speleothems' U–Th ages (dark blue circles, left *y* axis; Vaks et al.[14] and current work) and U–Pb ages (light blue circles, left *y* axis), the latter are given with percentages of radiogenic ^{206}Pb (red circles, right *y* axis). All age uncertainties are shown by 95% confidence error bars. The stratigraphic age of the layers dated from stalagmite SB-01112 increases from left to right.



Extended Data Fig. 9 | ^{238}U–^{234}U–^{206}Pb ages and concordance of ^{235}U–^{207}Pb ages for Ledyanaya Lenskaya Cave (based on Supplementary Table 1). Ages and age uncertainties are shown in black with the corresponding coloured error bar indicating the degree of concordance (^{235}U–^{207}Pb age as a percentage of the ^{238}U–^{234}U–^{206}Pb age; blue horizontal line indicates perfect concordance) of the ^{235}U–^{207}Pb age. Blue error bars indicate the analysis is concordant within error. Orange error bars denote slightly discordant analyses, where the apparent discordance is likely due to common Pb correction on ^{207}Pb. Red error bars denote analyses that are discordant with markedly low ^{235}U–^{207}Pb ages attributed to overcorrection of the common Pb due to a residual interference on ^{208}Pb. The numbers and bars in red show the percentage bias in the ^{238}U–^{234}U–^{206}Pb ages if the discordance of these samples is attributed to an interference on ^{208}Pb, based on the relative sensitivity of the ^{235}U–^{207}Pb and ^{238}U–^{234}U–^{206}Pb ages to the common Pb correction (Supplementary Table 1). Grey horizontal lines indicate where ages replicate in two or more speleothems. The biases in the ^{238}U–^{234}U–^{206}Pb ages attributed to the interference on ^{208}Pb do not change these ages outside of error, hence, the tendency of these ages to replicate well in other samples. The discordance arises almost entirely from the ^{235}U–^{207}Pb ages (which are not used), due to their vastly greater sensitivity to over-correction of the common Pb than the ^{238}U–^{234}U–^{206}Pb ages. Errors are 95% confidence. For each speleothem, the stratigraphic age of the dated layers increases from left to right.



Extended Data Fig. 10 | HYSPLIT-model-based seven-day back-trajectories of 12 snow events at the Ledyanaya Lenskaya Cave site. A–L, These snowfall events were accompanied by a substantial decrease in air temperature indicating that Arctic air was involved in the synoptic event. Six snow events occurred in October and first half of November 2012 (**A**–**F**), and other six events occurred in October and first half of November of 2013 (**G**–**L**). For each of 12 events, three back trajectories for elevations of 500, 1,500 and 5,000 m above sea level were calculated, with at least one of them originating in the Arctic Ocean in each case. The dates and times of the events are as follows (dd/mm/yyyy): 04/10/2012, 21:00 (**A**); 07/10/2012, 21:00 (**B**); 19/10/2012, 09:00 (**C**); 21/10/2012, 15:00 (**D**); 05/11/2012, 09:00 (**E**); 11/11/2012, 09:00 (**F**); 01/10/2013, 09:00 (**G**); 09/10/2013, 00:00 (**H**); 19/10/2013; 09:00 (**I**); 25/10/2013, 09:00 (**J**); 02/11/2013, 09:00 (**K**); 10/11/2013, 09:00 (**L**). Identification information: Job ID: 164576; Job Start: Tue 3 Sep, 12:19:42, UTC 2019; Meteorology: 0000Z, 1 Oct 2012 – GDAS1 (**A**); Job ID: 165119; Job Start: Tue 3 Sep, 12:46:49, UTC 2019; Meteorology: 0000Z, 1 Oct 2012 – GDAS1 (**B**); Job ID: 169518; Job Start: Sun 8 Sep 14:04:06, UTC 2019; Meteorology: 0000Z, 15 Oct 2012 – GDAS1 (**C**); Job ID: 169478; Job Start: Sun 8 Sep 13:58:40, UTC 2019; Meteorology: 0000Z, 22 Oct 2012 – GDAS1 (**D**); Job ID: 171083; Job Start: Sun 8 Sep 15:42:40, UTC 2019; Meteorology: 0000Z, 1Nov 2012 – GDAS1 (**E**); Job ID: 171195; Job Start: Sun 8 Sep 15:47:51, UTC 2019; Meteorology: 0000Z, 8 Nov 2012 – GDAS1 (**F**); Job ID: 187728; Job Start: Fri 30 Aug 12:07:03 UTC 2019; Meteorology: 0000Z, 1 Oct 2013 – GDAS1 (**G**); Job ID: 172040; Job Start: Sun 8 Sep 16:46:37 UTC 2019; Meteorology: 0000Z, 8 Oct 2013 – GDAS1 (**H**); Job ID: 172169; Job Start: Sun 8 Sep 16:54:44 UTC 2019; Meteorology: 0000Z, 15 Oct 2013 – GDAS1 (**I**); Job ID: 172846; Job Start: Sun 8 Sep 17:16:55 UTC 2019; Meteorology: 0000Z, 22 Oct 2013 – GDAS1 (**J**); Job ID: 173152; Job Start: Sun 8 Sep 17:24:37 UTC 2019; Meteorology: 0000Z, 1 Nov 2013 – GDAS1 (**K**); Job ID: 173482; Job Start: Sun 8 Sep 17:32:06 UTC 2019; Meteorology: 0000Z, 8 Nov 2013 – GDAS1 (**L**). Parameters equal for all calculations in Extended Data Fig. 10: Source, latitude 60.371000, longitude 116.946472; heights, 500 m, 1,500 m, 5,000 m above ground level (AGL); trajectory direction, backward; duration, 168 h; vertical motion calculation method, model vertical velocity.

Article

Long-term cyclic persistence in an experimental predator–prey system

https://doi.org/10.1038/s41586-019-1857-0

Received: 2 July 2019

Accepted: 18 November 2019

Published online: 18 December 2019

Bernd Blasius[1,2]*, Lars Rudolf[3], Guntram Weithoff[4,5], Ursula Gaedke[4,5] & Gregor F. Fussmann[6]

Predator–prey cycles rank among the most fundamental concepts in ecology, are predicted by the simplest ecological models and enable, theoretically, the indefinite persistence of predator and prey[1–4]. However, it remains an open question for how long cyclic dynamics can be self-sustained in real communities. Field observations have been restricted to a few cycle periods[5–8] and experimental studies indicate that oscillations may be short-lived without external stabilizing factors[9–19]. Here we performed microcosm experiments with a planktonic predator–prey system and repeatedly observed oscillatory time series of unprecedented length that persisted for up to around 50 cycles or approximately 300 predator generations. The dominant type of dynamics was characterized by regular, coherent oscillations with a nearly constant predator–prey phase difference. Despite constant experimental conditions, we also observed shorter episodes of irregular, non-coherent oscillations without any significant phase relationship. However, the predator–prey system showed a strong tendency to return to the dominant dynamical regime with a defined phase relationship. A mathematical model suggests that stochasticity is probably responsible for the reversible shift from coherent to non-coherent oscillations, a notion that was supported by experiments with external forcing by pulsed nutrient supply. Our findings empirically demonstrate the potential for infinite persistence of predator and prey populations in a cyclic dynamic regime that shows resilience in the presence of stochastic events.

Cyclic dynamics are one of the most notable phenomena in population biology and are known to occur in a large range of communities both in the wild[5–8,13,14] and the laboratory[9–12,15–18]. A number of mechanisms that cause populations to oscillate have previously been identified[3,4,20]; the most highly investigated, however, are the cyclic dynamics that arise from trophic interactions between populations of predator and prey organisms. The nearly century-old fascination with this type of dynamics is rooted in the predictions of simple mathematical models, which suggest that the predator and prey may coexist on recurring cyclic trajectories over indefinitely long periods of time[1,2]. Empirical data that support the sustained nature of predator–prey cycles in the field are difficult to come by because tractable populations typically have cycle periods of three to ten years[4,5,13]. An alternative data source are laboratory studies with fast-reproducing organisms, with which long-running experiments can be realized under strictly controlled conditions[9,10,12,16,21]. In a seminal experimental study, population numbers of weevils and their larval parasites were recorded for a time span of about 20 cycles but oscillations could not be sustained for the whole duration of the experiment[10]. This observation is paradigmatic for many studies, which showed that cycles in closed laboratory predator–prey systems tend to be short-lived, with populations either becoming extinct or reaching a stationary state.

These findings suggest that sustained, intrinsically generated predator–prey cycles might be a construct of ecological theory and that predator–prey systems require external drivers such as spatial structure, immigration or environmental perturbations to exhibit persistent cycles[9,11,12,17,22,23]. To explore the potential for long-term persistence of predator–prey cycles, we cultured freshwater organisms—planktonic rotifers together with their prey, unicellular green algae—in several independent chemostat trials under constant environmental conditions (Methods, experiments C1–C7). One of our experimental systems (Fig. 1a and Extended Data Fig. 1; C1 in Extended Data Table 1) persisted without any external stimuli in a homogeneous environment for over a year and displayed more than 50 cycles, which is equivalent to about 300 predator generations and at least as many prey generations. This result of the long-term persistence of the predator and prey in a cyclic dynamic regime was confirmed in four additional, replicated chemostat experiments with the same species (C2–C5 in Extended Data Table 1, Extended Data Fig. 2), and in two additional experiments (C6 and C7) with a different algal prey species (Extended Data Fig. 3). We applied power analysis and phase analysis (in particular, bivariate wavelet analysis[21,24–26]) (Methods) to quantify the cyclic succession and the distribution of phase differences in our measured time series. Wavelet analysis revealed statistically significant associations between the dynamics of predator and prey densities in our experiments and power analysis determined a mean period length of 6.7 days (C1–C5) (Fig. 1d and Extended Data Fig. 1d, f).

[1]Institute for Chemistry and Biology of the Marine Environment ICBM, University of Oldenburg, Oldenburg, Germany. [2]Helmholtz Institute for Functional Marine Biodiversity, University of Oldenburg (HIFMB), Oldenburg, Germany. [3]Institute of Physics and Astronomy, University of Potsdam, Potsdam, Germany. [4]Department of Ecology & Ecosystem Modelling, University of Potsdam, Potsdam, Germany. [5]Berlin-Brandenburg Institute for Advanced Biodiversity Research, Berlin, Germany. [6]Department of Biology, McGill University, Montréal, Québec, Canada. *e-mail: blasius@icbm.de

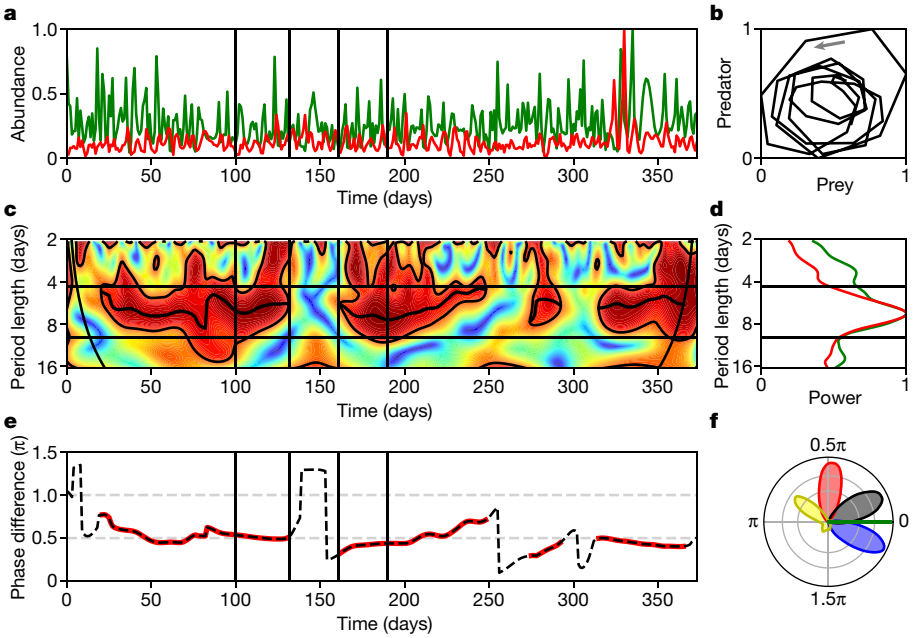


Fig. 1 | Phase analysis of a year-long, oscillatory predator–prey time series (experiment C1). a, Time series of the predator, *B. calyciflorus* (red) and its prey, *Monoraphidium minutum* (green); abundance normalized to a range of 0 to 1. **b**, Phase portrait of the bandpass-filtered predator–prey time series between days 90 and 135; the arrow indicates the direction of oscillation. **c**, Wavelet coherence (WCO) between predator and prey. Values range from 0 (blue) to 1 (red). The cone of influence is indicated by black thin curved lines. The black contour lines enclose significant areas (95% significance level, WCO > 0.83). The thick black line segments show the instantaneous oscillation period $\tilde{s}(t)$ of highest WCO within these areas (coherent oscillation regimes). **d**, Global wavelet power spectrum $P(s)$ for the predator (red) and the prey (green). **e**, Phase difference $\theta(t)$ between predator and prey; well-defined relations between both signals are marked in red. **f**, Circular distribution of the mutual phase difference to prey (indicated in green) for the predator (red), the predator egg ratio (blue), the number of eggs (black) and the dead animals (yellow). **c**, **d**, The two horizontal lines indicate the prefixed period band of $[s_1, s_2]$. **a**, **c**, **e**, The solid vertical lines indicate selected time spans of regular and irregular oscillations, which are enlarged in Fig. 2.

Transitions between different oscillation regimes

Although the experiments were not perturbed by external influences, we identified sharp transitions between two different dynamic regimes. For most of the time, we observed oscillations with a significantly coherent, nearly constant phase relationship (coherent oscillation regimes) (Methods), in which the predator densities followed the prey densities with a phase lag of about $\theta = 0.5\pi$ (equivalent to 90°), consistent with classical predator–prey theory. These periods with coherent oscillations were intersected by non-coherent oscillation regimes in which the well-defined phase relationship between the measured signals was lost, but re-initiated shortly after (Fig. 1c, e). For example, the time interval between days 100 and 131 in C1 (Fig. 2a, b) and between days 37 and 66 in C7 (Fig. 2c, d) showed sustained cycles with regular period lengths and a constant phase lag between prey and predator densities, which is also reflected in the anticlockwise motion in the prey–predator phase plane (Fig. 1b). This regime of coherent oscillations gave way to a non-coherent, more irregular time series (days 132–161 in C1 and days 67–97 in C7), in which the phase relationship between prey and predator was lost, although both populations continued to oscillate (Fig. 2). Finally, and without external intervention, the system displayed resilience and the characteristic predator–prey phase difference was re-established.

Combining the analyses of all five experiments with the same predator and prey species (C1–C5), we found that coherent oscillations were the dominant type of dynamics and comprised 66% of the total experimental time with a typical regime duration of 58 days (Extended Data Fig. 4 and Methods). In the remaining 34% of the experiments, predator and prey continued to oscillate, but without a well-defined phase relationship (typical duration of non-coherent regimes, 23 days) (Methods). Phase analysis of surrogate data revealed that spurious coherent oscillation regimes can arise also by chance, but with a significantly smaller duration (typical duration, 15 days) (Extended Data Fig. 4).

We performed additional analyses and experiments that were designed to provide mechanistic insights into the origin of the observed coherent oscillations and the intermittent breaks in the phase signature. First, we extended the specificity of the phase analysis by including predator life-history stages. Second, we formulated and analysed a mathematical model that accounts for the predator's stage structure. And third, we performed three additional chemostat experiments: two experiments with periodic changes in the nutrient concentration in the supply medium, to explore how the phase relations were influenced by external forcing (C8 and C9); and one experiment at dilution rates that, according to bifurcation theory[2,16] and our mathematical model, should cause the predator–prey system to cross over from oscillatory to equilibrium dynamics (C10).

Phase analysis of life-history stages

The predator in our system, the rotifer *Brachionus calyciflorus*[27], is a small freshwater metazoan that reproduces asexually in the chemostats (only females occur) and undergoes a stage-structured life cycle. Reproducing females carry 1–5 eggs, from which juveniles hatch and grow to adulthood without any larval stages; adults die shortly after the last egg has hatched. In addition to analysing algal and rotifer densities, we applied phase analysis to all discernible predator life-history stages: the number of eggs, the egg ratio (that is, the average number of eggs per female rotifer) and the number of dead animals. We found that these life-history parameters also exhibited persistent cycles that were coherently locked in phase with prey abundance (Fig. 3a). These phase differences were notably constant over all different experimental replicates (Extended Data Figs. 1–3). Averaging across all experiments (C1–C5), we obtained the following phase signature: the predator density followed the prey density with a phase difference of 93° ± 21°, the

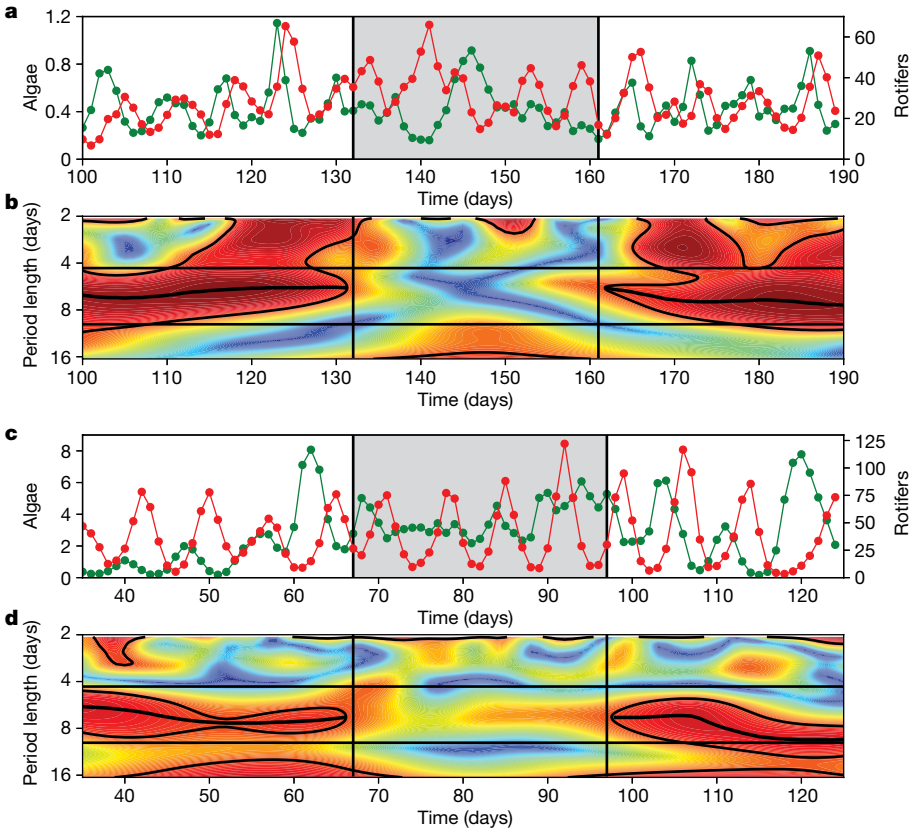


Fig. 2 | Resilience of phase relationship in experiments with two different prey species. a, Enlarged view of the densities of the predator (animals per millilitre, red) and prey (*M. minutum*, 10^6 cells per millilitre, green) in experiment C1 for a time window of 90 days. **b**, Enlarged view of their WCO during the same time interval (colours and lines as in Fig. 1c). **a**, **b**, Solid vertical lines border a time interval without significant wavelet coherence between predator and prey, separating three dynamic regimes: days 100–131, distinct predator–prey cycles with a constant phase shift of $\theta = \pi/2$; days 132–161, regime of non-coherent oscillations (grey background); days 162–190, spontaneous re-initiation of the regular phase-locked oscillation. **c**, **d**, As in **a**, **b**, but showing a time window of 90 days for experiment C7 with a different algal prey, *Chlorella vulgaris*. Non-coherent oscillations arise between days 67 and 97.

abundance of eggs was delayed to the abundance of prey with a phase difference of 27° ± 19°, whereas the egg ratio preceded the prey signal (−27° ± 16°); dead animals had a phase difference of −160° ± 68°. These mutual phase differences can be compactly visualized in a single circular histogram[21] (Figs. 1f and 3b). This 'phase signature' enables the straightforward interpretation of the temporal succession of the community during the cycles and thus provides a fingerprint for the cyclic community structure.

Mathematical model

We compared our experimental results with numerical simulations of a stage-structured predator–prey model that included the operative parameters of the chemostat (inflow nutrient concentration and dilution rate) and essential organismal variables (algae as well as rotifer eggs, juveniles, adults and dead animals) (Extended Data Fig. 5 and Methods). Numerical simulations revealed sustained oscillations with a period length of 6.8 days (experimental mean, 6.7 days). When we introduced stochasticity into the model, we observed similar breaks in the oscillation regimes as in experiments C1–C7, with random transitions between coherent and non-coherent oscillating regimes that were triggered by stochasticity alone. We were still able to recover a phase signature that aligned well with the experimental observations (predator, 90° ± 11°; eggs, 11° ± 13°; egg ratio, −21° ± 2°; dead rotifers, −150° ± 60°) (Fig. 3c, d and Extended Data Fig. 5).

Notably, both in the data and the model, the predator egg ratio preceded the phase of prey availability, which is not immediately obvious as egg production is directly connected to food availability. This phase lag between prey density and egg ratio disappeared in model simulations when the juvenile maturation rate was set to zero (Extended Data Fig. 6), revealing the effect of juveniles on the phase difference signature. Juvenile predators (which, as immature individuals, carry no eggs by definition) have a crucial influence as they lead to an early breakdown of the egg-ratio maximum, caused by the multitude of non-egg-bearing juveniles during the prey maximum, finally giving rise to a negative phase difference between egg ratio and prey (Fig. 3a, c and Extended Data Fig. 6a). Thus, as physiological information may be encoded in the phase difference distributions, extending phase analysis to life-history stages enabled deeper insights into the functioning of our system.

Additional experiments

In experiments C8 and C9, we externally forced the dynamics through pulsed supply of nutrients (Methods). We chose a forcing period (8.0 days) that was clearly distinguishable from the period of the unforced experiments (6.7 days) but sufficiently close to enable, in theory, predator and prey (that is, living populations with demographic constraints) to lock to the external signal. Indeed, we observed predator–prey oscillations that were locked to the external signal, both in chemostat experiments and model results (Fig. 3e, g and Extended Data Fig. 7). As expected, oscillations were persistently coherent, without the non-coherent breaks in the time series from the unforced experiments (C1–C7), although intermittent period doubling occurred in experiment C8 (Fig. 3e and Extended Data

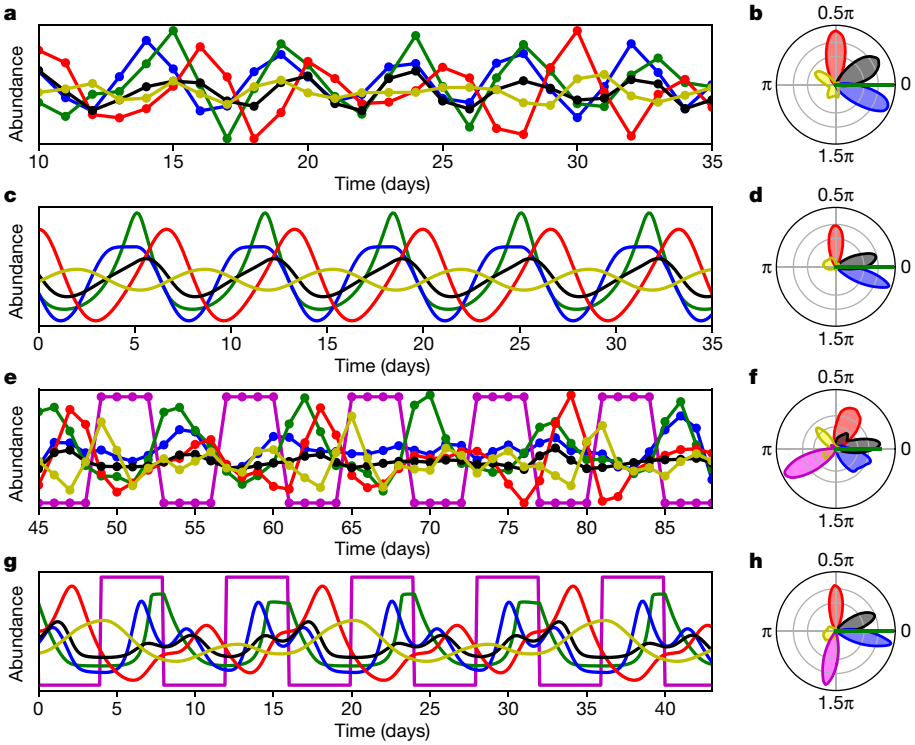


Fig. 3 | Dynamics and phase relationships in unforced versus externally forced systems. a, Time series of algae (green), rotifers (red), eggs (black), egg ratio (blue) and dead animals (yellow) in an unforced (that is, 'free-running' under constant conditions) experiment covering five oscillations (days 10–35 in C2). Data were filtered with a second order Butterworth bandpass filter and are shown in relative units. **b**, Phase signature aggregated over all experiments with the algal prey *M. minutum* under unforced conditions (C1–C5, including a total duration of 965 days), showing the circular distribution of the mutual phase differences to prey (colour code as in **a**). **c**, Simulated time series from the predator–prey model under constant conditions with all stochasticity removed (mean centred, relative units). **d**, Phase signature obtained from stochastic simulations of the predator–prey model reproduces the experimentally observed phase relationships and variances in **b**. **e**, As in **a**, but for an experiment with external forcing (days 45–88 of C8). The concentration of the external medium is shown in magenta. **f**, Phase signature aggregated over all experiments with external forcing (C8–C9, including a total duration of 323 days). **g**, **h**, As in **c**, **d**, but for simulations in an externally forced system.

Fig. 7c, d). Notably, externally forced and unforced experiments produced essentially the same phase signature (Fig. 3b, f), which was repeated also in model simulations (Fig. 3d, h). Thus, even in a dynamic regime that was strongly influenced by and locked to an external factor, phase analysis of different life-history stages was able to reveal the underlying predator–prey structure and to disentangle between intrinsic and external origins of cycles. This demonstrates that phase signature analysis can be a powerful diagnostic tool for the analysis of field data, for which it is often not known whether oscillations are the result of external factors or originate from species interactions.

Finally, we confirmed that the coherent oscillations observed in our experimental system indeed represent stable limit cycles, which arise from the destabilization of a stable equilibrium point. Oscillations in experiment C10 persisted only in a certain parameter range; if the dilution rate was increased, the oscillations died out (Extended Data Fig. 8). This result agrees with model predictions and previous experiments showing that increased dilution rates drive the system over a Hopf bifurcation[16] (Extended Data Fig. 9).

Conclusions

Persistent and coherent predator–prey oscillations are a potential dynamic regime enabling the prolonged coexistence of predator and prey populations, as predicted by fundamental theory. In live systems, stochasticity and other external perturbations can obscure or prevent predator–prey oscillations, and our simulations suggest that stochasticity is probably responsible for the reversible shifts from coherent to more-erratic oscillations that we observed in our highly controlled experiments. Four separate lines of evidence support a causal relationship between sustained cycles and predator–prey interactions. (1) Cyclic patterns dominated throughout the entire experiment and the system displayed resilience—that is, it returned to a well-defined phase relationship after intermittent non-coherent periods. (2) Planned experimental interventions enabled us to suppress cycles (C10) and to quench intermittent non-coherent oscillations (C8 and C9). (3) Data on the stage structure of the predator population enabled us to explicitly link characteristics of the predator–prey cycles to recurring changes in the demographic structure of the predator population. Therefore, our case for sustained predator–prey cycles is much stronger than it is possible for unstructured microbial populations, in which population densities are the only variables. (4) Finally, the experimental predator–prey cycles manifested as a regular phase-locked succession, a pattern that we were able to capture as a community phase signature. Our approach firmly establishes this phase signature—which, in a similar way, can be constructed for any oscillating system—as a robust measure for identifying species interactions, cyclic or seasonal species succession, and the regulatory dependencies that underlie community dynamics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1857-0.

1. Volterra, V. Fluctuations in the abundance of a species considered mathematically. *Nature* **118**, 558–560 (1926).
2. May, R. M. Limit cycles in predator–prey communities. *Science* **177**, 900–902 (1972).
3. Kendall, B. E. et al. Why do populations cycle? A synthesis of statistical and mechanistic modeling approaches. *Ecology* **80**, 1789–1805 (1999).
4. Berryman, A. A. (ed.) *Population Cycles: The Case for Trophic Interactions* (Oxford Univ. Press, 2002).
5. Elton, C. & Nicholson, M. The ten-year cycle in numbers of the lynx in Canada. *J. Anim. Ecol.* **11**, 215–244 (1942).
6. Blasius, B., Huppert, A. & Stone, L. Complex dynamics and phase synchronization in spatially extended ecological systems. *Nature* **399**, 354–359 (1999).
7. Gilg, O., Hanski, I. & Sittler, B. Cyclic dynamics in a simple vertebrate predator–prey community. *Science* **302**, 866–868 (2003).
8. Ryabov, A. B., de Roos, A. M., Meyer, B., Kawaguchi, S. & Blasius, B. Competition-induced starvation drives large-scale population cycles in Antarctic krill. *Nat. Ecol. Evol.* **1**, 0177 (2017).
9. Gause, G. F., Smaragdova, N. P. & Witt, A. A. Further studies of interaction between predators and prey. *J. Anim. Ecol.* **5**, 1–18 (1936).
10. Utida, S. Cyclic fluctuations of population density intrinsic to the host–parasite system. *Ecology* **38**, 442–449 (1957).
11. Huffaker, C. B. Experimental studies on predation: dispersion factors and predator–prey oscillations. *Hilgardia* **27**, 343–383 (1958).
12. Luckinbill, L. S. Coexistence in laboratory populations of *Paramecium aurelia* and its predator *Didinium nasutum*. *Ecology* **54**, 1320–1327 (1973).
13. Krebs, C. J. et al. Impact of food and predation on the snowshoe hare cycle. *Science* **269**, 1112–1115 (1995).
14. Hudson, P. J., Dobson, A. P. & Newborn, D. Prevention of population cycles by parasite removal. *Science* **282**, 2256–2258 (1998).
15. McCauley, E., Nisbet, R. M., Murdoch, W. W., de Roos, A. M. & Gurney, W. S. C. Large-amplitude cycles of *Daphnia* and its algal prey in enriched environments. *Nature* **402**, 653–656 (1999).
16. Fussmann, G. F., Ellner, S. P., Shertzer, K. W. & Hairston, N. G. Jr. Crossing the Hopf bifurcation in a live predator–prey system. *Science* **290**, 1358–1360 (2000).
17. Ellner, S. P. et al. Habitat structure and population persistence in an experimental community. *Nature* **412**, 538–543 (2001).
18. Becks, L., Hilker, F. M., Malchow, H., Jürgens, K. & Arndt, H. Experimental demonstration of chaos in a microbial food web. *Nature* **435**, 1226–1229 (2005).
19. Massie, T. M., Blasius, B., Weithoff, G., Gaedke, U. & Fussmann, G. F. Cycles, phase synchronization, and entrainment in single-species phytoplankton populations. *Proc. Natl Acad. Sci. USA* **107**, 4236–4241 (2010).
20. Radchuk, V., Ims, R. A. & Andreassen, H. P. From individuals to population cycles: the role of extrinsic and intrinsic factors in rodent populations. *Ecology* **97**, 720–732 (2016).
21. Benincà, E., Jöhnk, K. D., Heerkloss, R. & Huisman, J. Coupled predator–prey oscillations in a chaotic food web. *Ecol. Lett.* **12**, 1367–1378 (2009).
22. Holyoak, M. & Lawler, S. P. Persistence of an extinction-prone predator–prey interaction through metapopulation dynamics. *Ecology* **77**, 1867–1879 (1996).
23. Stenseth, N. C. et al. The effect of climatic forcing on population synchrony and genetic structuring of the Canadian lynx. *Proc. Natl Acad. Sci. USA* **101**, 6056–6061 (2004).
24. Torrence, C. & Compo, G. P. A practical guide to wavelet analysis. *Bull. Am. Meteorol. Soc.* **79**, 61–78 (1998).
25. Cazelles, B. et al. Wavelet analysis of ecological time series. *Oecologia* **156**, 287–304 (2008).
26. Keitt, T. H. Coherent ecological dynamics induced by large-scale disturbance. *Nature* **454**, 331–334 (2008).
27. Paraskevopoulou, S., Tiedemann, R. & Weithoff, G. Differential response to heat stress among evolutionary lineages of an aquatic invertebrate species complex. *Biol. Lett.* **14**, 20180498 (2018).
28. Guillard, R. R. L. & Lorenzen, C. J. Yellow-green algae with chlorophyllide *c*. *J. Phycol.* **8**, 10–14 (1972).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Experiments

We ran chemostat experiments with parthenogenetic rotifers (*B. calyciflorus* sensu strictu[27], a small metazoan freshwater zooplankton species) as predators and unicellular algae (*M. minutum* or *C. vulgaris*) as prey under constant temperature (23 °C) and permanent illumination. Experiments were conducted (mostly consecutively) in the same well-maintained climate chamber. We carefully searched for possible sources of external perturbations, but all measured parameters (for example, temperature and irradiance) remained constant throughout the experimental runs. For inoculation, we used stock cultures originally raised from a single individual and added *B. calyciflorus* 10 days after the algae had reached a biomass that enabled rotifer growth. The chemostats (experimental volume, 0.8 l) had a constant inflow of sterile medium with a rate of 0.55 per day. We used a modified WC medium[28] with a reduced nitrogen concentration of 80 µmol l^{-1} nitrate as the limiting resource. In the experiments with external forcing, we periodically changed the concentration in the external medium to between 0 and 160 µmol l^{-1} nitrate with a period length of 8 days. We took daily subsamples of 8 ml to determine the abundance of both predator and prey. The algal abundance was analysed with an electronic particle counter (CASY, Schärfe). The rotifers were counted using an inverted microscope at 100-fold magnification. For the rotifers, we recorded the total number of rotifers and of asexually produced subitaneous eggs and the dead animals. No males or sexually reproducing females were observed throughout the experiment. The egg ratio (that is, the total number of eggs divided by the total number of animals in the population) is a measure of the reproductive status of the population.

Bivariate wavelet analysis

Because the measured signals showed large variability, we applied phase analysis[6], which relies on the fact that the regulatory dependence between state variables is often encoded in their phase relationship, whereas the amplitudes may be highly erratic and uncorrelated[6]. The wavelet method[24] is a particularly powerful tool for extracting optimally resolved phase information from ecological time series[25,26] and enables us to quantify transient associations between two non-stationary signals[24,29–32].

Continuous wavelet transformation

The continuous wavelet transformation[24] (CWT) of a signal $x(t)$ is calculated as the convolution of the signal with a localized complex-valued wavelet function $\psi(\tau)$ centred at time t and dilated by the scale parameter s

$$W_x(s,t) = \frac{1}{\sqrt{s}} \int \mathrm{d}t' x(t') \psi^* \left(\frac{t'-t}{s} \right)$$

Here, the asterisk denotes the complex conjugate. After decomposing the complex CWT into amplitude and phase $W_x(s,t) = |W_x(s,t)| e^{i\phi_x(s,t)}$ intuitively it can be understood as the extent $|W_x(s,t)|$ to which the signal $x(t)$ at local time t resembles an oscillation with period length s and phase $\phi_x(s,t)$.

Throughout our analysis, we used the Morlet wavelet $\psi(\tau) = \pi^{-1/4} e^{i\omega_0 \tau} e^{-\tau^2/2}$ with $\omega_0 = 6$ so that the scale parameter s can simply be interpreted as the period length of the oscillation[24]. We always took s from 100 uniform steps per octave on a logarithmic scale. Edge effects were minimized by applying zero padding to the time series to the next highest integer power of two. We always indicate the cone of influence, which marks the zone in the CWT that is affected by edge effects[24].

Wavelet power spectrum

Although the wavelet transform is complex, the real-valued local wavelet power spectrum (WPS) of a signal $x(t)$ at time t and scale s can be defined as

$$W_{xx}(s,t) = \langle W_x(s,t)\ W_x^*(s,t) \rangle$$

in which '⟨…⟩' denotes a smoothing operator in both scale and time (see 'Parametrization and smoothing'). The global WPS $P_x(s)$ is defined as the time average of the local WPS and measures the averaged variance of the signal $x(t)$ at scale s. We used the maximum of $P_x(s)$ to estimate the mean oscillation period of the signal.

Wavelet cross-spectrum

To quantify the statistical relationship between two non-stationary signals $x(t)$ and $y(t)$, the wavelet cross-spectrum (WCS) is defined as the smoothed product of the corresponding wavelet transforms[24,29–31]

$$W_{x,y}(s,t) = \left\langle W_x(s,t)\ W_y^*(s,t) \right\rangle$$

The WCS is complex-valued. After decomposition into amplitude and phase $W_{x,y}(s,t) = A_{x,y}(s,t)\ e^{i\Delta\phi_{x,y}(s,t)}$, the phase describes the de-lay $\Delta\phi_{x,y}(s,t) = \phi_x(s,t) - \phi_y(s,t)$ between the two signals at time t and scale s, whereas the amplitude $A_{x,y}(s,t)$ expresses the covarying power of the two processes.

Wavelet coherence

The WCS is not specific, because it exhibits high values when either a true covariance between the two signals exists or when one of the spectra exhibits a high value. To solve this problem, a normalized measure is given by the WCO[29–31], which is defined as the amplitude of the WCS divided by the square roots of the two local WPS

$$\mathrm{WCO}_{x,y}(s,t) = \frac{\left| W_{x,y}(s,t) \right|}{W_{x,x}(s,t)^{1/2}\ W_{y,y}(s,t)^{1/2}}$$

The WCO estimates the relationship between the two signals at time t and scale s, normalized into the range $0 \le \mathrm{WCO} \le 1$. As the WCO is real-valued, it cannot be used to extract the delay between two signals, in contrast to the WCS.

Dominant phase difference

The phase difference $\Delta\phi_{x,y}(s,t)$ between two signals $x(t)$ and $y(t)$, as obtained from the WCS, depends on the scale parameter s. Assume now that the two signals oscillate with a common, but possibly time-varying, period length. To obtain their dominant phase difference, we define for every time instance t the dominant scale $\tilde{s}(t)$ as the scale parameter that has the maximal WCO at time t over all scales in a pre-fixed band $[s_1, s_2]$:

$$\tilde{s}(t) = \mathrm{argmax}_{s_1 < s < s_2} \mathrm{WCO}(s,t)$$

These values $\tilde{s}(t)$ constitute a line of the strongest co-oscillating component of both signals in the timescale plane[32]. Inserting this line in the WCS yields the dominant phase difference $\theta_{x,y}(t)$ between the two signals at every time instance

$$\theta_{x,y}(t) = \phi_x\ (\tilde{s}(t), t) - \phi_y(\tilde{s}(t), t)$$

Significance testing

To assess the statistical significance of the WCO, we generated $n = 1{,}000$ independent realizations of surrogate data by time-scrambling the

empirical signals (that is, sampling 1,000 random values with replacement from the measured predator and prey time series). The resulting surrogate data retain the frequency histogram of the original data but lose all temporal correlations. Using Monte Carlo simulations, we found that for our chosen parameterization (see 'Parametrization and smoothing') two signals are significantly correlated at time t and scale s on a 95% critical level if WCO(s, t) = 0.83.

Coherent oscillation regime

We defined a maximal contiguous time interval for which the dominant scale $\tilde{s}(t)$ is within a statistically significant area in the timescale plane—that is, being inside the cone of influence and having a significant WCO($\tilde{s}(t)$, t) > 0.83—as a coherent oscillation regime. The remaining time intervals constitute the non-coherent oscillation regimes. Each signal on its own may well oscillate in a regular fashion even within a non-coherent oscillation regime, but for a pair of signals the wavelet analysis guarantees a significantly correlated oscillation—and therefore a well-defined phase relationship—only in a coherent oscillation regime.

To characterize the typical duration $\tilde{T}$ of coherent oscillations in an ensemble of signals, we first measure the length T_i of each coherent oscillation regime i = 1 ... N that occurs in the ensemble and then calculate the expected oscillation regime duration, $\tilde{T}$, of a randomly chosen time instance taken from a coherent oscillation regime $\tilde{T} = \sum_{i=1}^{N} T_i^2 / \sum_{i=1}^{N} T_i$. We found that this measure gives a better description of the typical time to observe a coherent oscillation interval than the median or the mean $\overline{T} = \sum_{i=1}^{N} T_i / N$ of the length distribution.

Circular phase distribution

To characterize the phase relationship between a pair of covarying signals, we calculated the dominant phase differences θ_k for all time instances within a coherent oscillation regime. We used circular statistics[33] to characterize the resulting distribution of phase differences θ_k. For this we calculated the complex order parameter z, which is defined as the average of the complex numbers $e^{i\theta_k}$

$$z = Re^{i\overline{\theta}} = \frac{1}{N}\sum_{k=1}^{N} e^{i\theta_k}$$

Here the phase angle $\overline{\theta}$ of z indicates the circular mean of the phase differences, and the amplitude R of z gives the circular standard deviation of the distribution as $\sqrt{-2\ln R}$.

Given the sampling frequency of 1 day^{-1}, the Nyquist frequency is 0.5 day^{-1}, which for an average period length of 6.7 days roughly corresponds to a phase resolution of 27°. Averaging phase differences over many cycles enabled us to achieve a much better phase resolution, yielding typically rather smooth distributions of phase differences. We verified that these smooth phase distributions are no artefacts, by comparing phase distributions obtained from our highly resolved numerical simulations (see 'Numerical simulation') with those obtained from time series sampled from these simulations once per day, which yielded very similar results.

Phase signature

To characterize the mutual phase relationship of not only two but also an ensemble of time series (for example, total densities of predator and prey and of measured life history stages), we calculated the distribution of phase differences of all time series with respect to one selected reference time series. The combined plot of all resulting circular distributions in one common polar histogram defines our 'phase signature', providing a fingerprint for the cyclic succession of the corresponding state variables.

Parametrization and smoothing

Following a previously published study[31], we performed smoothing in time by convolution with a normalized real Morlet wavelet $\exp[-\tau^2/(2s^2)]/(s\sqrt{2\pi})$. Smoothing in scale direction was obtained by averaging over a constant-length (boxcar) window over 0.6 octaves—the scale decorrelation length for the Morlet wavelet—on a logarithmic scale[31]. Similarly, we took the upper and lower scale band borders s_1 and s_2 as ± 0.6 octaves from the peak in the global (that is, time-averaged) WCS. For a better visualization, circular histograms (bin width of π/100) were smoothed by a moving average (window width 0.15π, the above calculated phase resolution).

We verified that the resulting phase distributions remain largely unchanged for different parameterizations of the wavelet analysis, in particular for different wavelet functions, different values of the wavelet parameter ω_0, different smoothing parameters and scale band borders s_1 and s_2.

Numerical simulation

We developed a mathematical model to describe a stage-structured predator–prey community in a chemostat[34,35], which closely follows our experimental setup. The model includes as state variables the concentration of nitrogen N, phytoplankton P, and three life stages of the predator, namely eggs E, juveniles J and adults A. Additionally the model keeps track of the dead animals D in the chemostat. We assume a constant time delay θ for egg development, a constant time delay τ for juvenile maturation, and the same constant mortality rate for juveniles and adults (deliberately keeping the model simple, as intake-related regulations[34,35] did not improve the agreement between simulation results and data). We assume that the phytoplankton population is unstructured. All variables are measured in µmol N l^{-1}.

The model is written as the following system of time-delayed ordinary differential equations:

$$\begin{aligned}
\dot{N} &= \delta N_{in} - F_P(N)P - \delta N \\
\dot{P} &= F_P(N)P - F_B(P)B/\varepsilon - \delta P \\
\dot{E} &= R_E - R_J - \delta E \\
\dot{J} &= R_J - R_A - (m+\delta)J \\
\dot{A} &= \beta R_A - (m+\delta)A \\
\dot{D} &= m(J+A) - \delta D
\end{aligned}$$

Here, β is the adult/juvenile mass ratio (the juvenile/egg mass ratio is assumed to be 1) and the recruitment rates are

$$\begin{aligned}
R_E(t) &= F_B(P(t))\,A(t) \\
R_J(t) &= R_E(t-\theta)\,\exp(-\delta\theta) \\
R_A(t) &= R_J(t-\tau)\,\exp(-\delta\tau)
\end{aligned}$$

Algal nutrient uptake $F_P(N)$ is modelled as a Monod function and predator recruitment as a type-3 functional response with Hill coefficient κ

$$F_P(N) = r_P \frac{N}{K_P + N}$$
$$F_P(P) = r_B \frac{P^\kappa}{K_B^\kappa + P^\kappa}$$

The total predator density is given by $B(t) = \beta J(t) + A(t)$ and the egg ratio is $\beta E(t)/B(t)$. As state variables are given in units of nitrogen, juveniles and eggs here need to be scaled by the adult/juvenile mass ratio β.

The model is solved numerically by an Euler scheme with a time step of h = 0.001 days. We included environmental stochasticity by taking the growth rates r_P and r_B as autocorrelated random numbers by multiplying them in every time step i with a factor of $1 + s_i$.

Autocorrelation (that is, red noise) is implemented as $s_{i+1} = as_i + \xi_i$ (first-order autoregressive AR(1) process[36]) with autocorrelation parameter a and independent and identically distributed Gaussian numbers ξ_i with mean 0 and variance $\sigma^2(1-a^2)$. For wavelet analysis and for representation in figures, the simulated states are sampled with a rate of 1 per day, in accordance with the experiments. To capture measurement error and additional stochastic influences, we replaced the sampled rotifer state variables, $E(t)$, $J(t)$, $A(t)$ and $D(t)$, by random values drawn from a Poisson distribution with an expectation value equal to that of the state variable. Finally, algal and rotifer state variables were rescaled to individuals per volume using the constants v_P, and v_B, respectively.

Parameter values were taken as follows. Mass ratio adult/juveniles, $\beta = 5$; dilution rate, $\delta = 0.55$ per day; nitrogen concentration in the external medium, $N_{in} = 80$ µmol l^{-1}; phytoplankton maximal growth rate, $r_P = 3.3$ per day; phytoplankton half-saturation constant: $k_P = 4.3$ µmol l^{-1}; rotifer maximal egg-recruitment rate, $r_B = 2.25$ per day; rotifer half-saturation constant, $k_B = 15$ µmol l^{-1}; Hill coefficient of the functional response, $\kappa = 1.25$; predator assimilation efficiency, $\varepsilon = 0.25$; rotifer mortality rate, $m = 0.15$ per day; egg development time, $\theta = 0.6$ days; juvenile maturation time, $\tau = 1.8$ days; autocorrelation parameter, $a = 0.9$; noise strength, $\sigma = 0.5$; nitrogen content per adult female *Brachionus*, $v_A = 0.57 \times 10^{-3}$ µmol N per female; nitrogen content per algal cell, $v_P = 28 \times 10^{-9}$ µmol N per cell.

Environmental parameters (δ, N_{in}) were taken according to the experimental set-up. Parameters for uptake and growth (r_P, k_P, r_B, k_B and ε) were taken according to a previous study[16]. The rotifer mortality rate (m) was increased compared with this previous study[16], as we did not include the loss of the reproducing rotifer cohort into senescent animals. The mass ratio of adults to juveniles (β) was taken from the literature[37]. Duration of development stages (θ, τ) was adjusted to be within the previously reported range[38]. We used a type-3 functional response[39,40]. Only the parameter values of population dynamics (θ, τ, κ and m) and stochasticity (noise levels and noise colour) were adjusted by hand within ecologically reasonable ranges to simultaneously optimally fit the comprehensive experimental data. As criteria we used the observed range and variability of the state variables, existence of oscillatory dynamics with correct period length, phase signatures (that is, phase relationships between all state variables, including the egg ratio), power and wavelet spectra, and the typical duration of coherent oscillatory regimes—both in the unforced and forced scenarios.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

Experimental data are publicly available on Figshare (https://doi.org/10.6084/m9.figshare.10045976.v1).

Code availability

Data analysis and the model were implemented in the language Julia v.1.1[41]. The source code for the wavelet analysis is publicly available on GitHub (https://github.com/berndblasius/WaveletAnalysis).

29. Torrence, C. & Webster, P. J. Interdecadal changes in the ENSO–monsoon system. *J. Clim.* **12**, 2679–2690 (1999).
30. Maraun, D. & Kurths, J. Cross wavelet analysis: significance testing and pitfalls. *Nonlin. Processes Geophys.* **11**, 505–514 (2004).
31. Si, B. C. Spatial scaling analyses of soil physical properties: a review of spectral and wavelet methods. *Vadose Zone J.* **7**, 547–562 (2008).
32. Bandrivskyy, A., Bernjak, A., McClintock, P. & Stefanovska, A. Wavelet phase coherence analysis: application to skin temperature and blood flow. *Cardiovasc. Eng.* **4**, 89–93 (2004).
33. Batschelet, E. *Circular Statistics in Biology* (Academic, 1981).
34. McCauley, E., Nisbet, R. M., de Roos, A. M., Murdoch, W. W. & Gurney, W. S. C. Structured population models of herbivorous zooplankton. *Ecol. Monogr.* **66**, 479–501 (1996).
35. de Roos, A. M. & Persson, L. *Population and Community Ecology of Ontogenetic Development 59* (Princeton Univ. Press, 2013).
36. Massie, T. M., Weithoff, G., Kuckländer, N., Gaedke, U. & Blasius, B. Enhanced Moran effect by spatial variation in environmental autocorrelation. *Nat. Commun.* **6**, 5993 (2015).
37. Mourelatos, S., Pourriot, R. & Rougier, C. Taux de filtration du rotifère *Brachionus calyciflorus*: comparaison des méthodes de mesure; influence de l'âge. *Vie Milieu* **40**, 39–43 (1990).
38. Halbach, U. Einfluß der Temperatur auf die Populationsdynamik des planktischen Rädertieres *Brachionus calyciflorus* Pallas. *Oecologia* **4**, 176–207 (1970).
39. Fussmann, G. E., Weithoff, G. & Yoshida, T. A direct, experimental test of resource vs. consumer dependence. *Ecology* **86**, 2924–2930 (2005).
40. Seifert, L. I. et al. Heated relations: temperature-mediated shifts in consumption across trophic levels. *PLoS ONE* **9**, e95046 (2014).
41. Bezanson, J., Edelman, A., Karpinski, S. & Shah, V. B. Julia: a fresh approach to numerical computing. *SIAM Rev.* **59**, 65–98 (2017).

Acknowledgements This work was supported by German VW-Stiftung. We thank A. Bandrivskyy, T. Massie, E. Denzin and all other technical staff of the Department of Ecology & Ecosystem Modelling; and R. Ceulemans, C. Feenders, A. Jamieson-Lane, A. Ryabov and E. van Velzen for comments on the manuscript.

Author contributions B.B., G.F.F., G.W. and U.G. designed the research, L.R., G.F.F. and G.W. performed the experiments, B.B. and L.R. performed the numerical simulations and data analysis, B.B. and G.F.F. wrote the paper. All authors discussed the results and commented on the manuscript.

Competing interests The authors declare no competing interests.

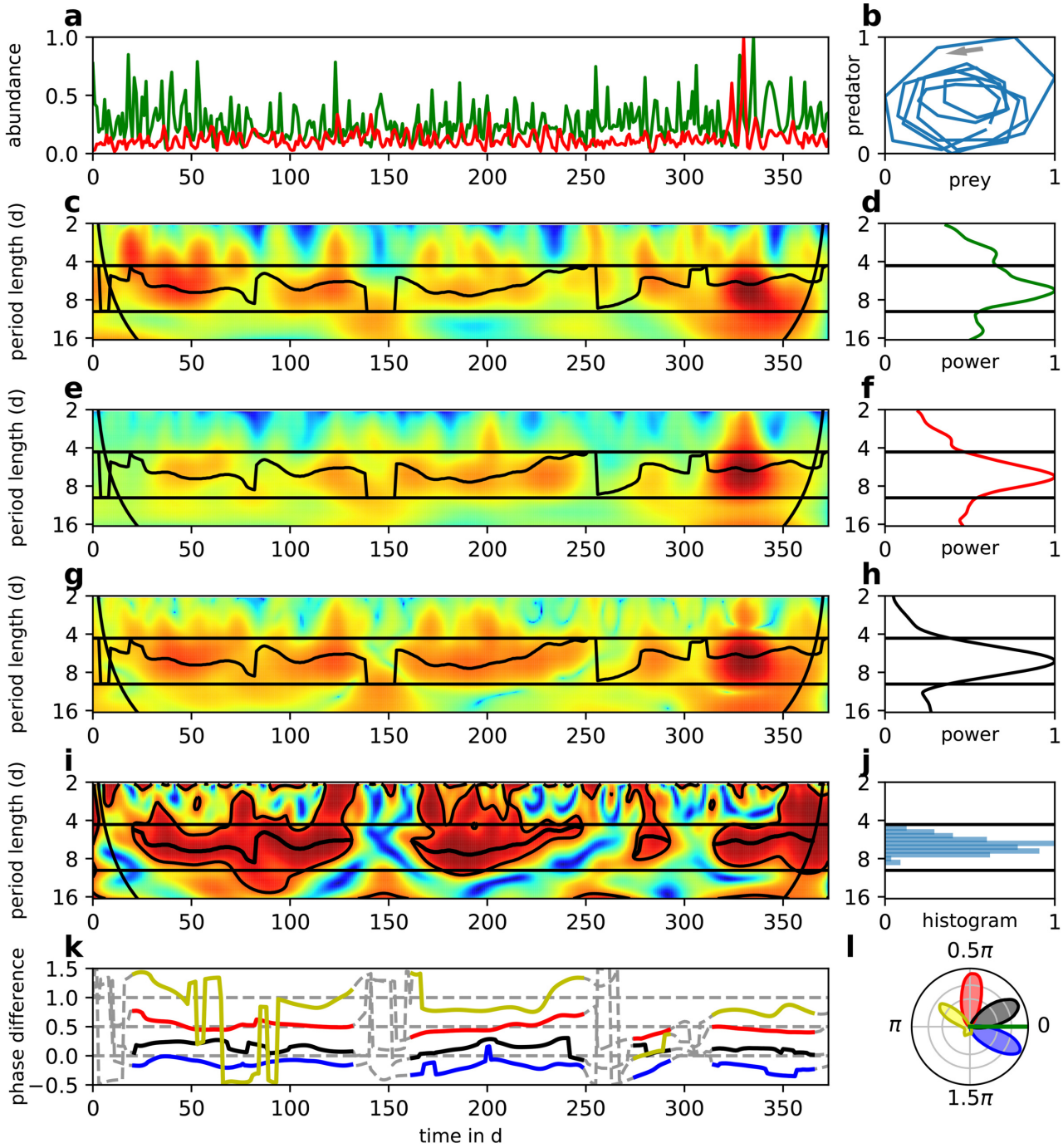
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1857-0.

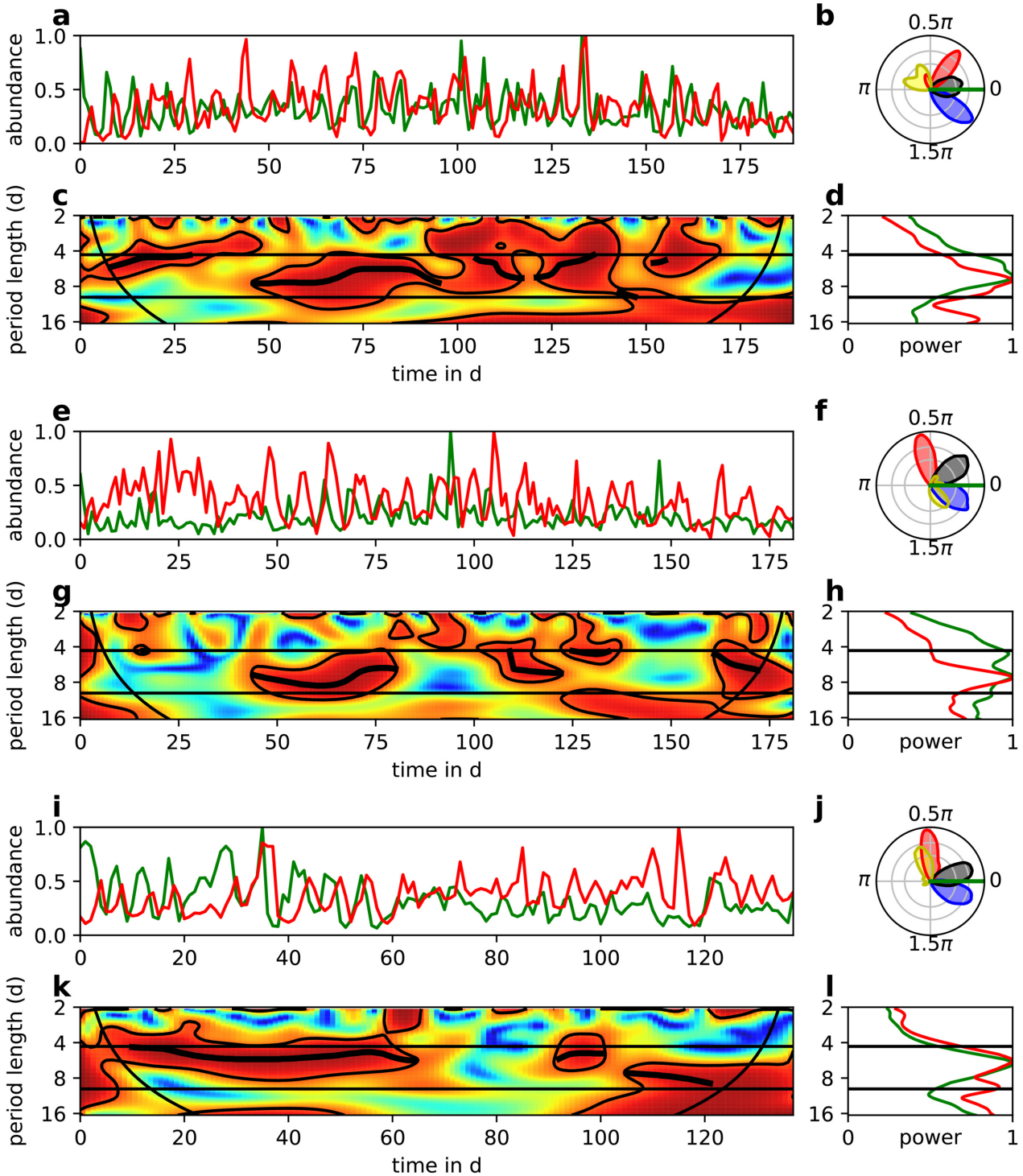
Correspondence and requests for materials should be addressed to B.B.

Peer review information *Nature* thanks Alan Hastings and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

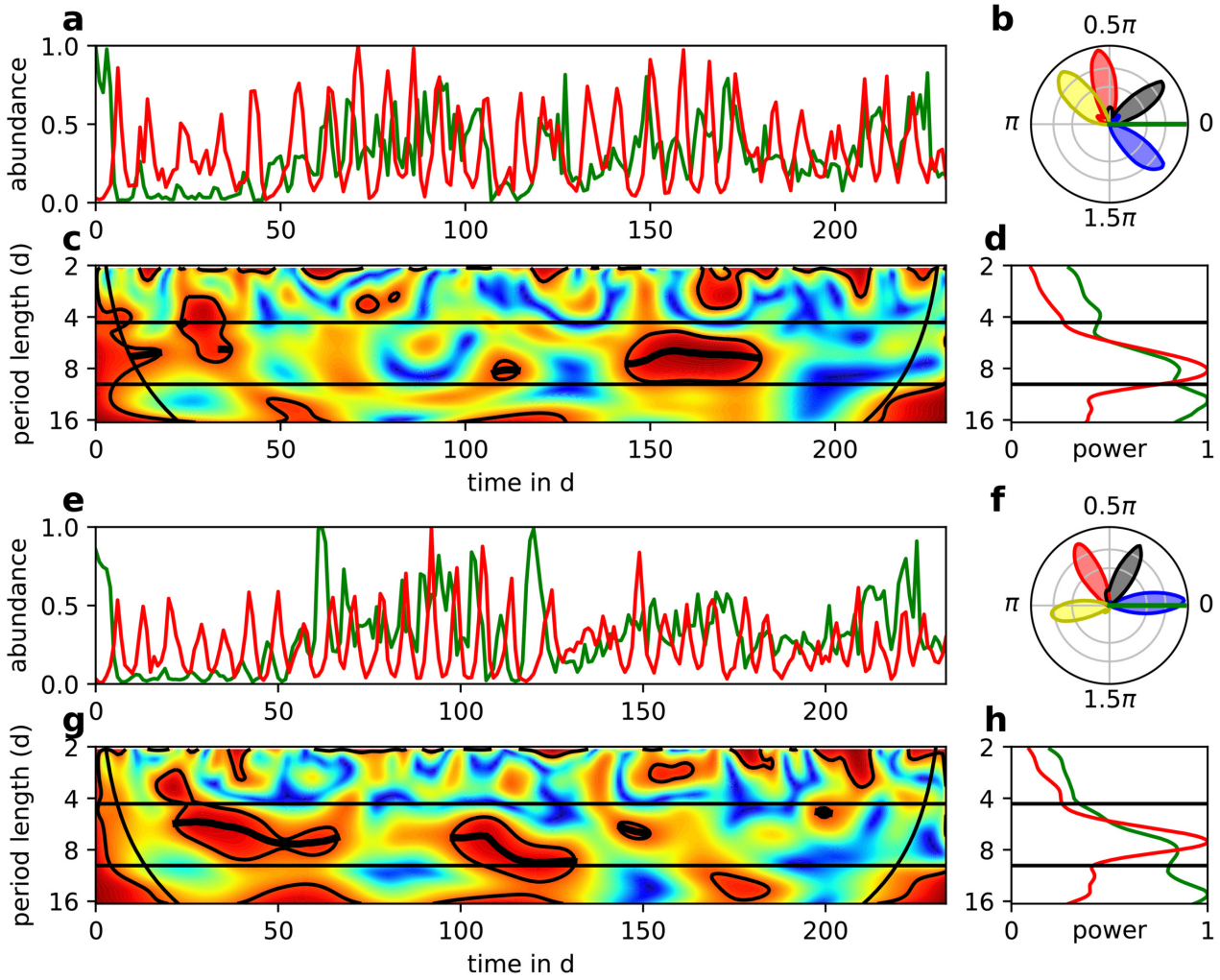
Reprints and permissions information is available at http://www.nature.com/reprints.



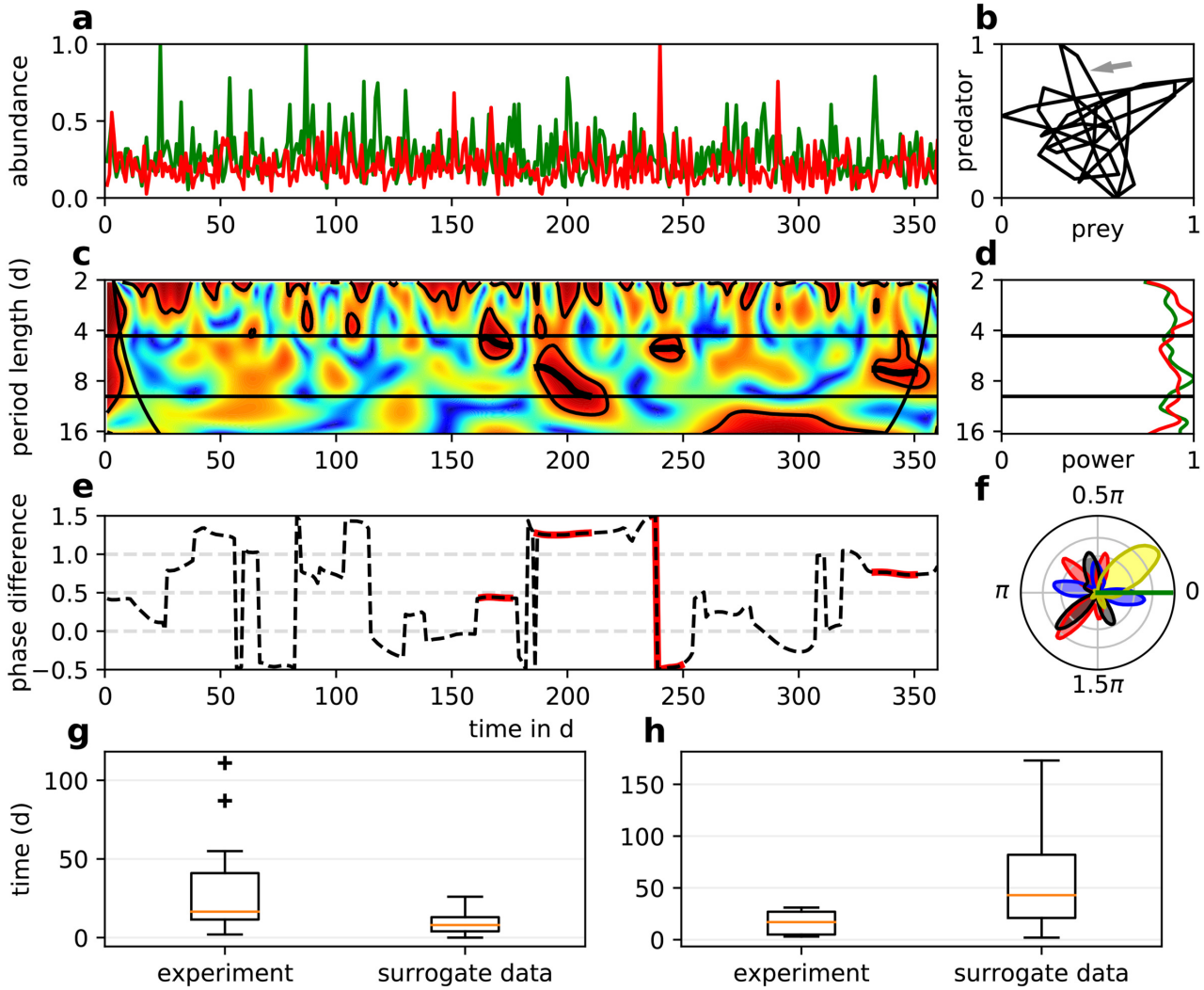
Extended Data Fig. 1 | Detailed dynamics and phase relationship in experiment C1. a, Time series of the predator, *B. calyciflorus* (red), and its prey, *M. minutum* (green), (normalized to the range of 0 and 1). **b**, Phase portrait of the bandpass-filtered predator–prey time series between days 90 and 130. The arrow indicates the direction of oscillation. **c**, Local WPS for the prey, ranging from 0 (blue) to maximum (red), and the instantaneous oscillation period $\tilde{s}(t)$ (black line) of the highest WCO within the prefixed period length band (horizontal black lines). The cone of influence is also indicated (black). **d**, Global WPS $P(s)$ for the prey (green). Horizontal lines as in **c**. **e**, **f**, As in **c**, **d**, but for the predator. **g**, WCS between predator and prey, colour-coding and other lines as in **c**. **h**, Amplitude of the global WCS for the predator and the prey. **i**, WCO between predator and prey, ranging from 0 (blue) to 1 (red). Significant areas (WCO > 0.83) are enclosed by thin solid lines; the thick black line segments show the instantaneous oscillation period $\tilde{s}(t)$ in these segments. **j**, Histogram of period lengths within coherent oscillation regions as detected in **i**. **k**, Phase difference $\theta(t)$ to prey for the predator (red), the predator egg ratio (blue), the number of eggs (black) and dead animals (yellow). Significant relations between respective signals are marked as a solid line. **l**, Circular distribution of the mutual phase difference to prey (indicated in green) for the predator (red), the predator egg ratio (blue), the number of eggs (black) and dead animals (yellow).



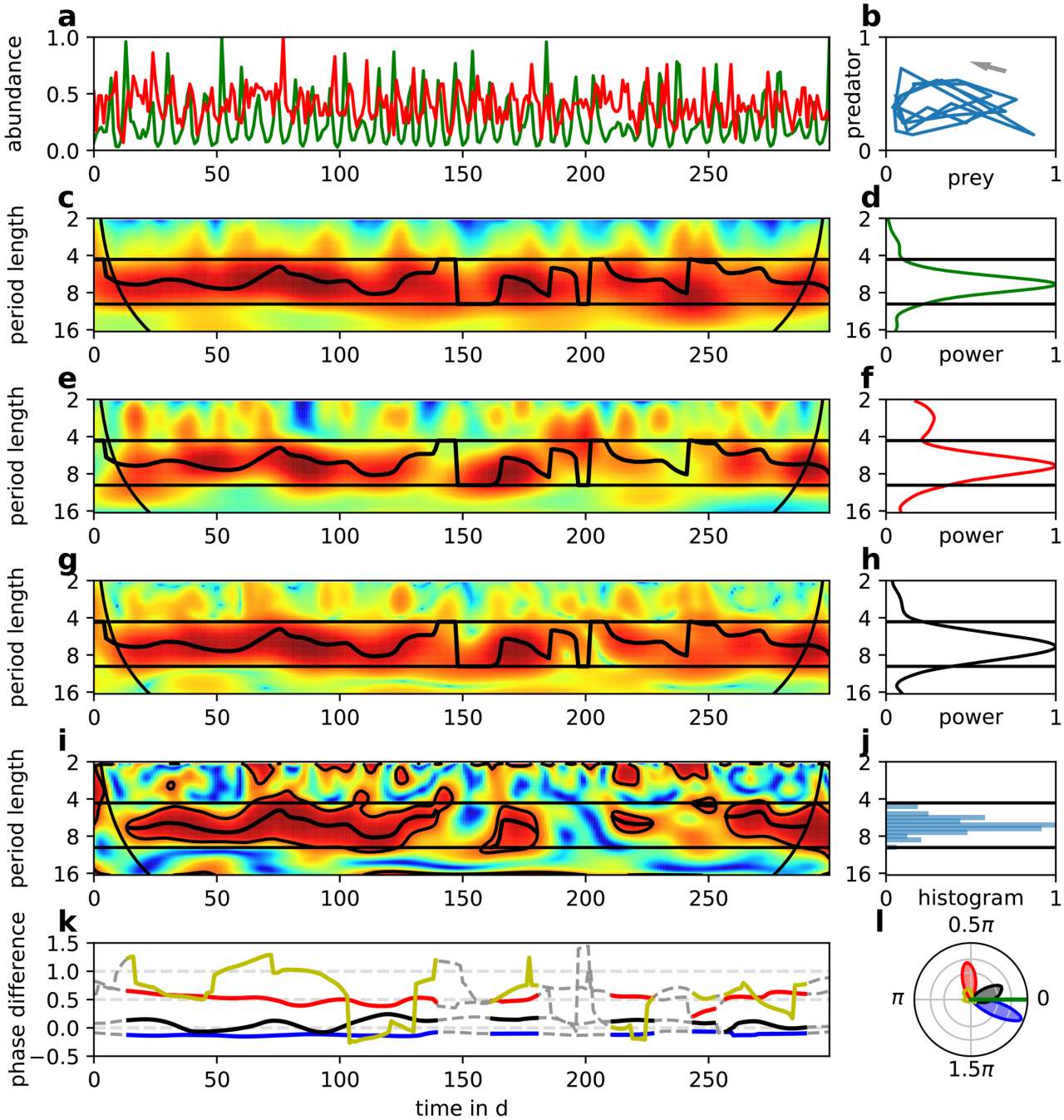
Extended Data Fig. 2 | Dynamics and phase relationships in three further experimental time series in a constant environment. a–d, Analysis of experiment C2. **a**, Predator–prey time series. **b**, Phase signature. **c**, WCO. **d**, Global WPS. **e–h**, As in **a–d**, for experiment C3. **i–l**, As in **a–d**, for experiment C4. Details and colours as in Fig. 1.



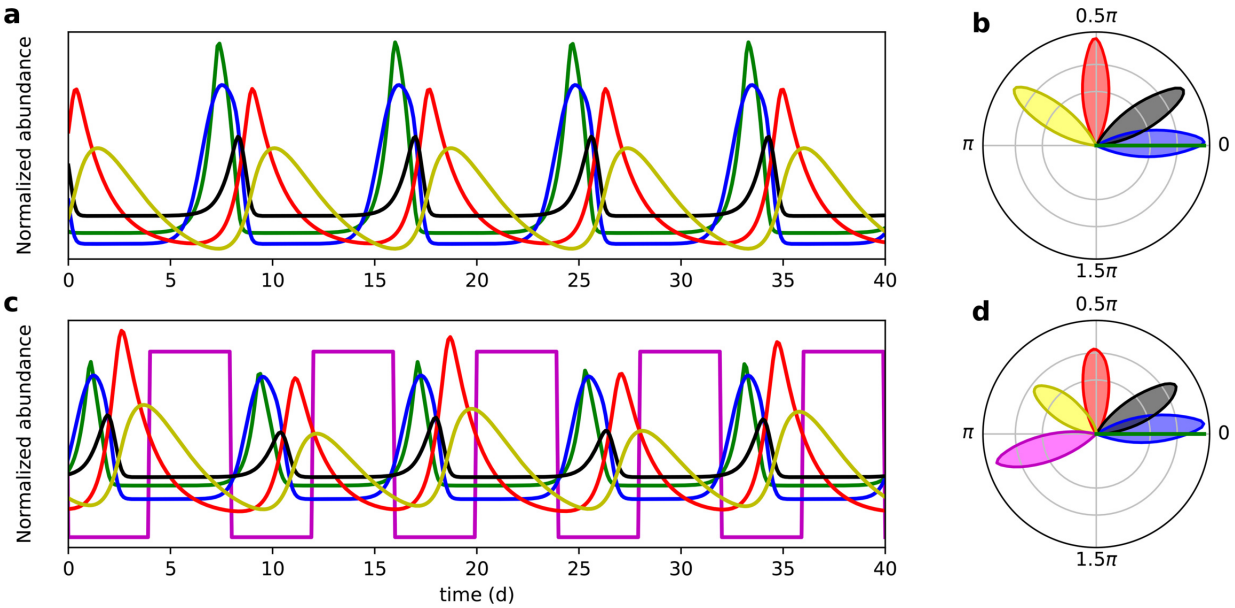
Extended Data Fig. 3 | Dynamics and phase relationships in measured time series with a different algal species (*C. vulgaris*). a–d, Analysis of experiment C6. **a**, Predator–prey time series. **b**, Phase signature. **c**, WCO. **d**, Global WPS. **e–h**, As in **a–d**, for experiment C7. Details and colours as in Fig. 1.



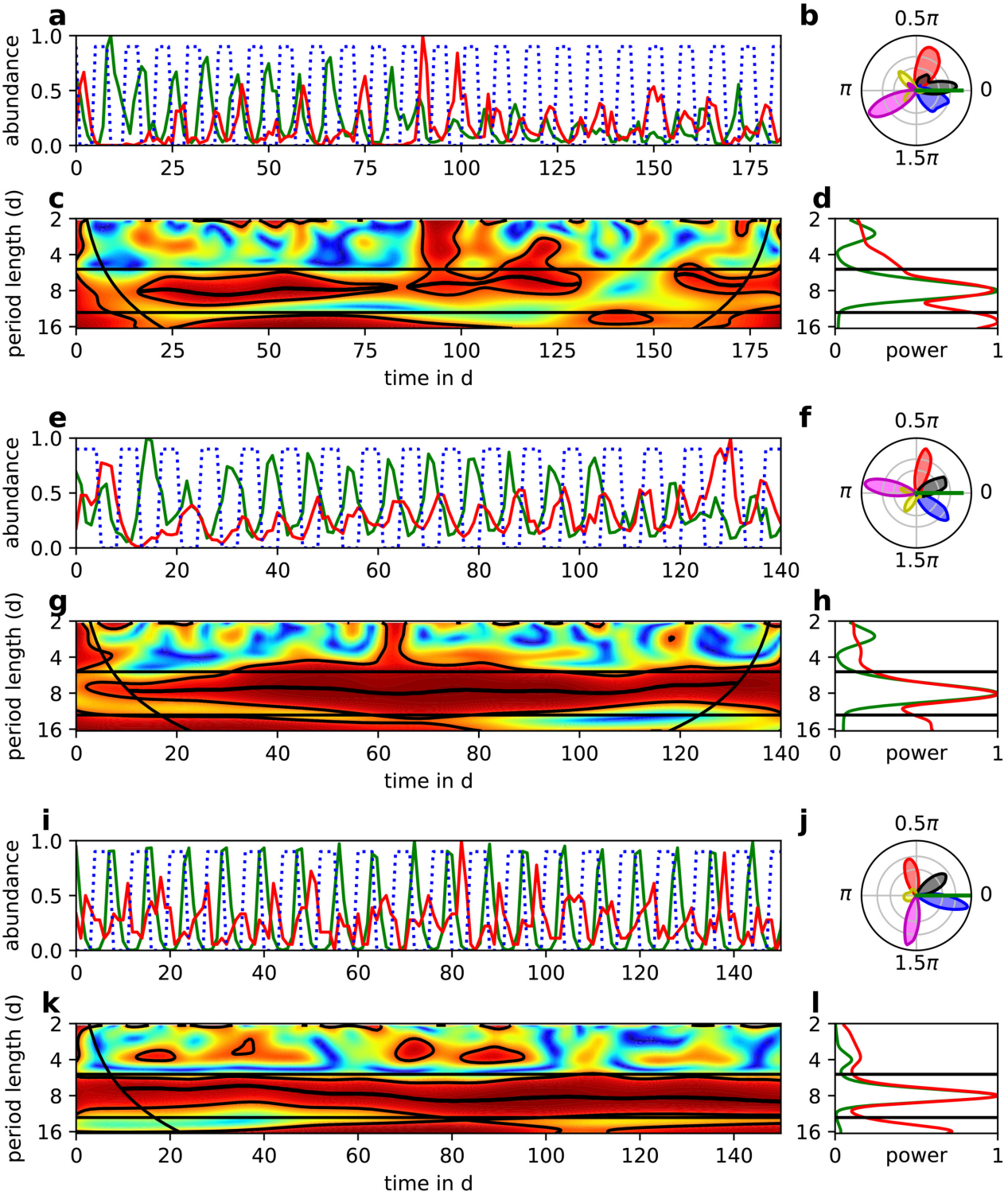
Extended Data Fig. 4 | Dynamics and phase relationship in surrogate data and length distribution of coherent oscillation regimes. a–**f**, As in Fig. 1, but for surrogate data, obtained by time scrambling—that is, randomly drawing 360 values with replacement from the predator and prey time series in experiment C1—yielding a time series without temporal correlation (white noise) but with the same abundance distribution as the experimental data. The figure shows a typically outcome. Occasionally, for short bursts the randomized predator and prey time series oscillate with a common frequency (**a**), yielding areas of high coherence (**c**; WCO > 0.83; spurious coherent oscillation regimes) with random phase differences (**e**). Whereas most time intervals with coherent predator–prey cycles in the surrogate data have a short duration, in this example (by chance) there is one regime (days 187–209) with a duration of 23 days. **g**, **h**, Box plots of oscillation regime length. Boxes show the interquartile range, orange lines indicate the median, whiskers range from the lowest to the highest data point within 1.5× the interquartile range and markers indicate outliers. **g**, Length distribution of coherent oscillation regimes for the experimental data under free-running conditions (C1–C5) and for a large ensemble of surrogate data. This shows that time intervals with coherent oscillations are significantly longer in the experiments ($n = 20$; median, 16 days; mean ± s.d., 29 ± 28 days) than in the surrogate data ($n = 13{,}719$; median, 8 days; mean ± s.d., 9 ± 7 days). The two outliers in the left box plot correspond to the two long coherent oscillation regimes of 111 days and 85 days in C1, which (even though they are outliers in the length distribution) effectively dominate the dynamic behaviour in C1. Consequently, we calculate the typical duration $\tilde{T}$ of coherent oscillations as the expected regime duration for a randomly chosen time instance taken from a coherent oscillation regime (Methods) ($\tilde{T} = 58$ days in C1–C5; $\tilde{T} = 15$ days in the surrogate data). **h**, As in **g**, but for durations of non-coherent oscillation regimes, showing that time intervals of non-coherent oscillations are significantly shorter in the experiments ($n = 15$; median, 17 days; mean ± s.d., 16 ± 11 days; typical duration $\tilde{T} = 23$ days) than in the surrogate data ($n = 12{,}719$; median, 43 days, mean ± s.d., 60 ± 11 days; typical duration $\tilde{T} = 113$ days).



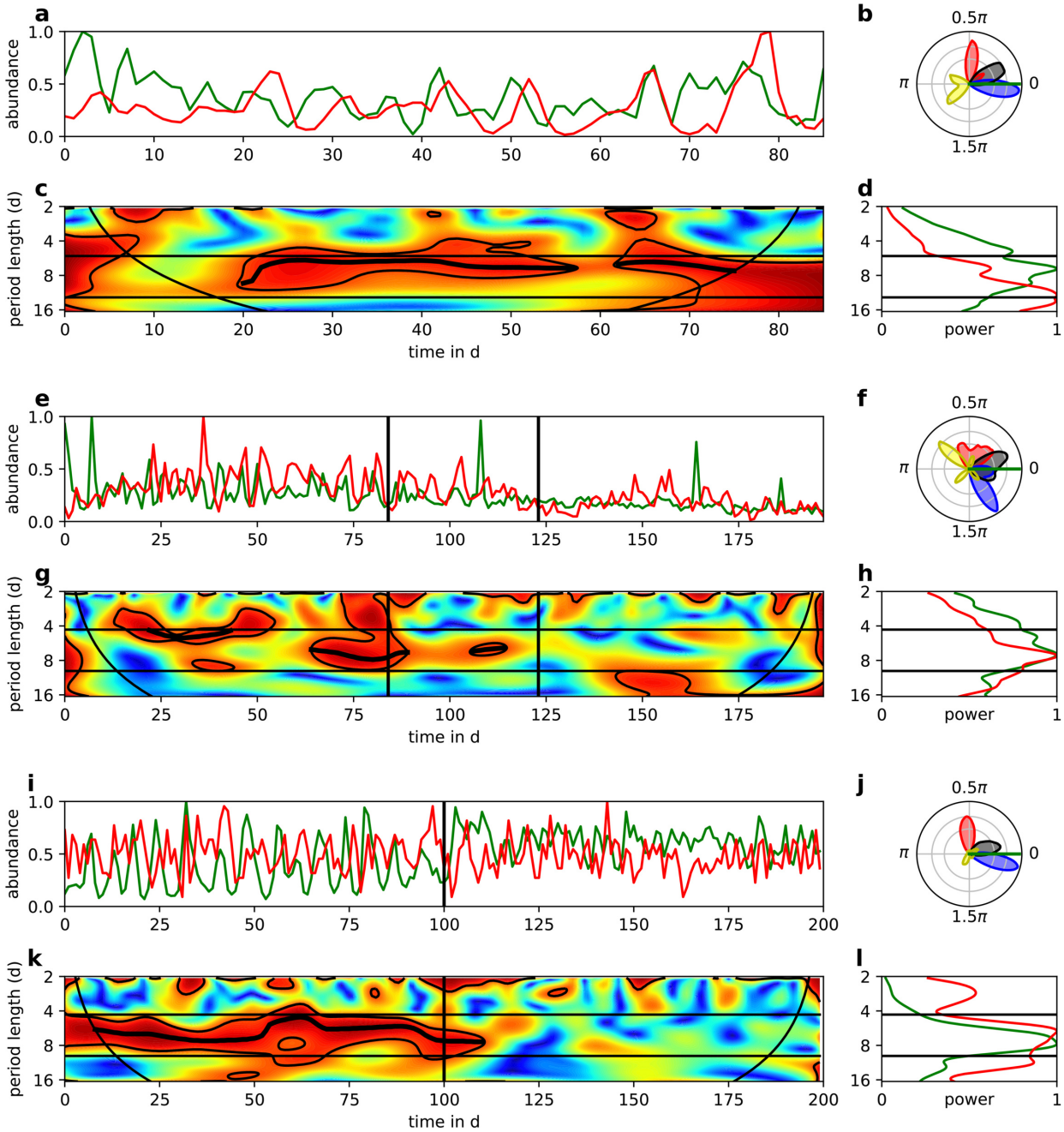
Extended Data Fig. 5 | Detailed dynamics and phase relationships in a modelled time series. a–l, As in Extended Data Fig. 1, but for a simulated time series of the stochastic time-delayed model in a constant environment. The simulation time was 300 days. A typical model outcome is shown. In this example, we obtained five coherent oscillation regimes (indicated as thick lines in **i**, **k**). Similar to the experimental systems (C1–C7), these regimes are interspersed by short non-coherent oscillation regimes without a significant phase relationship.



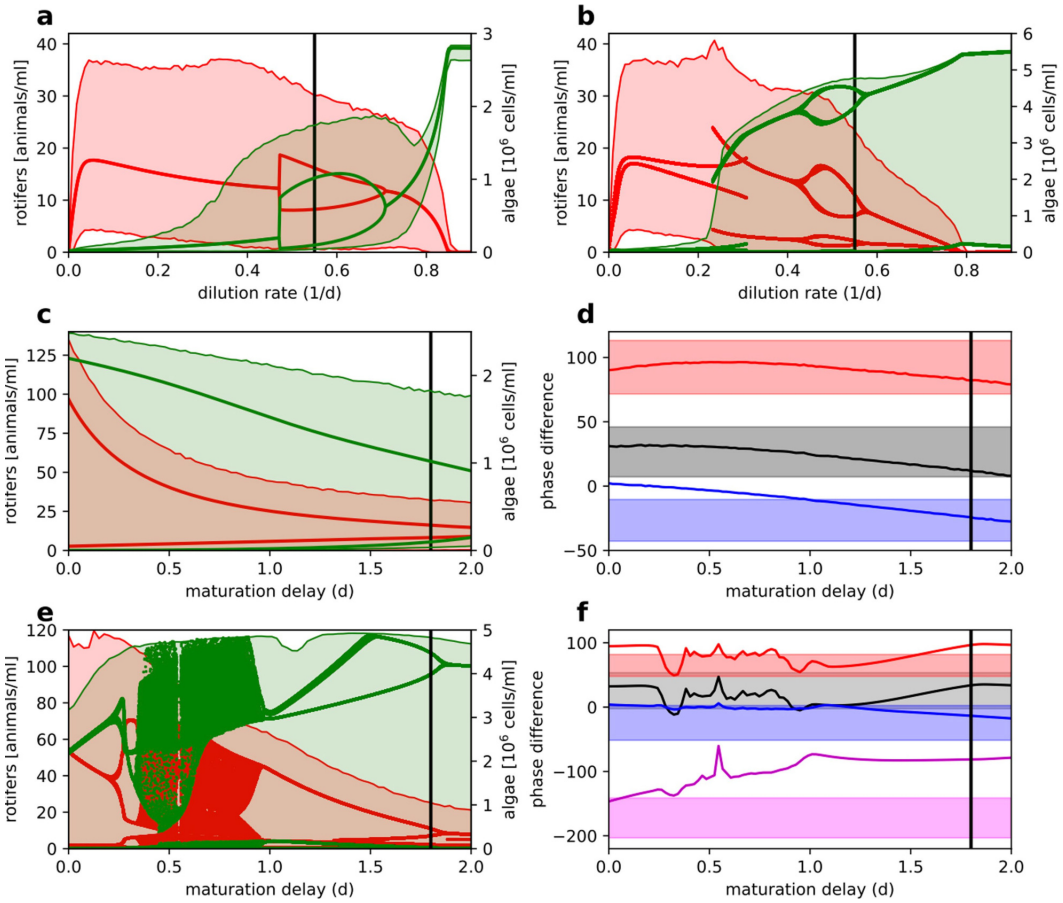
Extended Data Fig. 6 | Dynamics and phase relationships in a model without juvenile maturation delay. a, Simulated time series of algae (green), rotifers (red), eggs (black), egg ratio (blue) and dead animals (yellow) in a model without juvenile maturation delay, $\tau = 0$, under constant conditions with all stochasticity removed (mean centred, relative units). **b**, Phase signature obtained from stochastic simulations (free-running conditions, $\tau = 0$). **c**, **d**, As in **a**, **b**, but for an externally driven system (external nutrient concentration shown in magenta). The figure shows that in a scenario without juvenile maturation delay ($\tau = 0$), the phase of the egg ratio coincides with that of the algae, both in a free-running and in an externally driven model. This is in contrast to the experimental observations (Fig. 3, rows 1 and 3) and simulations with non-vanishing juvenile maturation delay ($\tau = 1.8$ days) (Fig. 3 rows 2 and 4, Extended Data Fig. 9d, f), in which the egg ratio is significantly preceding the algae signal. Thus, the phase signature helps to disentangle life-history mechanisms, as—by comparison of observed and simulated phase signatures—a delay of juvenile maturation is essential to attain the experimentally observed phase relations.



Extended Data Fig. 7 | Dynamics and phase relationships in externally driven systems. a–**d**, Analysis of experiment C8. **a**, Predator–prey time series. **b**, Phase signature. **c**, WCO. **d**, Global WPS. **e**–**h**, As in **a**–**d**, for experiment C9. **i**–**l**, As in **a**–**d**, for simulation results of an externally driven system. Details and colours as in Fig. 1. In **a**, **e**, **i**, the blue dotted line shows the input concentration in the external medium in normalized units. In **b**, **f**, **j**, the phase difference distribution between the external medium and the prey is shown in magenta.



Extended Data Fig. 8 | Dynamics and phase relationships in experiments and simulations with changed dilution rate and press perturbations. a–**d**, Analysis of experiment C5, which has an increased dilution rate ($\delta = 0.66$ per day). **a**, Predator–prey time series. **b**, Phase signature. **c**, WCO. **d**, Global WPS. **e**–**h**, As in **a**–**d**, for experiment C10, which undergoes press perturbations (vertical black lines) at day 84 (dilution rate was increased from $\delta = 0.55$ per day to $\delta = 1.2$ per day) and day 123 (dilution rate increased further to $\delta = 1.35$ per day). **i**–**l**, As in **a**–**d**, with simulation results of a system undergoing a press perturbation in which the dilution rate was increased to $\delta = 0.75$ per day at day 100 (vertical black line in **i**, **k**). Under conditions of the increased dilution rate, the oscillations are suppressed (Extended Data Fig. 9a). Details and colours as in Fig. 1.



Extended Data Fig. 9 | Bifurcation diagrams. a, Simulated values of predator (red) and prey (green) for different values of the dilution rate. Thick solid lines show maximal and minimal values obtained in the unforced model without stochasticity. Shaded areas show range of values in the stochastic model. Sustained predator–prey oscillations appear for $0.47 < \delta < 0.71$ per day. For large values of the dilution rate, $\delta > 0.87$ per day, the predator goes extinct. **b**, As in **a**, but for the externally forced system. The model shows oscillations for all parameters without predator extinction $\delta < 0.8$ per day, it exhibits a bistability regime (small- versus large-amplitude oscillations) for $0.23 < \delta < 0.31$ per day and a period-2 oscillation regime for $0.42 < \delta < 0.57$ per day. **c**, **d**, Bifurcation diagrams of the unforced model as a function of the juvenile maturation delay τ. **c**, As in **a**, but for the different bifurcation parameter. **d**, Phase difference to the algae for the predator (red), eggs (black) and the egg ratio (blue). The shaded areas indicate the range of phase differences (±1 s.d.) from the experiments in a constant environment (C1–C5). For large values of τ, the simulated phase signature aligns with the experimental data, whereas for small values of τ, the simulated phase difference of the egg ratio approaches zero. **e**, **f**, As in **c**, **d**, but for the externally forced model. **e**, For small delay times, $0.4 < \tau < 1$ day, the model exhibits a chaotic regime. **f**, Additionally the phase difference between algae and external nutrients is shown (magenta). Vertical solid lines show the actually used parameter values $\delta = 0.55$ per day and $\tau = 1.8$ days.

Extended Data Table 1 | Summary of all experiments

Code	Experimental conditions	Algal species	Duration (days)	N_i (μmol l^{-1})	δ (day^{-1})
C 1	constant environment	*M. minutum*	373	80	0.55
C 2	constant environment	*M. minutum*	189	80	0.55
C 3	constant environment	*M. minutum*	181	80	0.55
C 4	constant environment	*M. minutum*	137	80	0.55
C 5	constant environment	*M. minutum*	85	80	0.66
C 6	constant environment	*C. vulgaris*	230	160	0.55
C 7	constant environment	*C. vulgaris*	233	160	0.55
C 8	externally forced	*M. minutum*	183	0 / 160	0.55
C 9	externally forced	*M. minutum*	140	0 / 160	0.55
C 10	press perturbation	*M. minutum*	197	80	0.55/1.2/1.33

A series of ten chemostat experiments was performed, constituting a total of 1,948 measurement days (corresponding to 5.3 years of measurement) and covering 3 scenarios. (1) Constant environmental conditions (C1–C7, 1,428 measurement days). This scenario consisted of 4 trials with the alga *M. minutum* (C1–C4), 1 trial with *M. minutum* and an increased dilution rate of δ = 0.66 per day (C5), and 2 trials with the alga *C. vulgaris* and a doubled concentration of the external medium N_{in} = 160 μmol l^{-1} (C6 and C7). (2) Externally forced conditions (C8 and C9, 323 measurement days). The concentration of the external medium was periodically changed between N_{in} = 160 μmol l^{-1} and N_{in} = 0 μmol l^{-1} with a period length of 8 days. (3) Press perturbation (C10, 197 measurement days). The dilution rate was increased from the standard value δ = 0.55 per day in two steps to δ = 1.2 per day at day 84 and δ = 1.33 per day at day 123, suppressing the predator–prey oscillations.

nature research

Corresponding author(s): Bernd Blasius

Last updated by author(s): Oct 25, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Measured data were first transfered to Excel-files (Versions 9-14) and subsequently archieved as csv-files.
Data analysis	Data analysis and the model were implemented in the language Julia 1.1

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Experimental data were made publicly available on Figshare – doi.org/10.6084/m9.figshare.10045976.v1. Data analysis and the model were implemented in the language Julia 1.1 (ref. 41). The source code for the wavelet analysis was made publicly available on https.//github.com/berndblasius/WaveletAnalysis.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The aim of the study was to generate and analyse experimental long term data. Thus, the experimental design differs from the typical (multi-)factorial design. From our time series, the relevant parameters were extracted. We had four replicates for the core long term experiments. Treatment factors were nutrient supply (constant vs. pulsed), type of algae (Monoraphidium vs Chlorella) and flow rate (shift from standard flow rate to high rates). These were run in duplicate or triplicate, which is common for such long-term experimental studies (see Extended Data Table 1). Each run represents an individual independent chemostat vessel. Each vessel contained 800 ml culture medium. All experiments were run under stable conditions at 23 °C and permanent illumination.
Research sample	The experimental algae were Monoraphidium minutum (SAG-243-1, Göttingen, Germany) and Chlorella vulgaris (UTEX 26, Texas, USA). Stock cultures were kept under non-limiting conditions to avoid synchronization of the cells. Animals were Brachionus calyciflorus, strain IGB, grown in stock cultures with the experimental algae under the same conditions. All species are cosmopolitan species occuring in various types of freshwaters. During the experiments, a daily sample of 8 ml was taken, representing 1 % of the total experimental volume. This sample size is large enough to quantify the number of algae and animals (see below) and small enough in order not to change the flow rate.
Sampling strategy	The sample volume was constant throughout the study and chosen according to the expected number of algae/rotifers within the sample. Each experimental vessel contained up to several thousands of rotifers and up to several billions of algae. Since turbulent mixing prevents spatial heterogeneity both species are randomly distributed. A sample of 8 ml was taken (5 ml for the analysis of animals, 3 ml for the analysis of algae). The maximum abundance of the animals is higher than 100/ml so that a 5 ml sample is representative. As for the algae, the abundance is much higher (mio/ml), a smaller sub-sample was analysed using an electronic particle counter (CASY, Schärfe, Germany). Three sub-samples of 400µl were analysed and the mean was calculated. The coefficient of variation was on average 0.015.
Data collection	Data were collected daily, mostly by E. Denzin, or other staff members, guided by G. Weithoff and G. Fussmann. Each person who participated in sampling was individually trained to avoid sampling bias. The same holds for the quantitative analysis of algae and rotifers. The raw data from the electronic particle counter were archived on a computer and also transferred to Excel-files (Versions 9 - 14). The data from the microscopic quantification of the animals were first recorded on paper in a laboratory folder and then transferred to Excel-files (Versions 9 - 14). The correct transfer was doublechecked in order to avoid any error through the transfer process. From these files, csv-files were created and used further onwards.
Timing and spatial scale	The first experiments started in April 2001 and lasted until May 2009; data acquisition (counting) and analyses extended until 2019. Depending on human and spatial resources up to four chemostats were run simultaneously, often only two. The sampling frequency of daily samples was chosen in accordance to the growth rate of the experimental organisms. Daily samples allow for capturing the dynamic population behavior of the organisms and the subsequent analysis of population oscillations.
Data exclusions	Experimental runs were discarded from the data set when technical problems (leakage of tubes, blackout of pumps or illumination etc.) occurred before the data set was sufficiently long to extract the key parameters for this study.
Reproducibility	The reproducibility is part of the study, demonstrating that the findings were recurrent in several independent experimental runs.
Randomization	Treatments were set up independently and placed randomly in the climate chamber or different experimental runs, which were run subsequently, were run at the same spot.
Blinding	Blinding was neither possible, nor necessary. The correct type of medium, and flow rate had to be visible to the experimentalists.

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Mapping child growth failure across low- and middle-income countries

https://doi.org/10.1038/s41586-019-1878-8

Received: 16 November 2018

Accepted: 14 November 2019

Published online: 8 January 2020

Open access

Local Burden of Disease Child Growth Failure Collaborators*

Childhood malnutrition is associated with high morbidity and mortality globally[1]. Undernourished children are more likely to experience cognitive, physical, and metabolic developmental impairments that can lead to later cardiovascular disease, reduced intellectual ability and school attainment, and reduced economic productivity in adulthood[2]. Child growth failure (CGF), expressed as stunting, wasting, and underweight in children under five years of age (0–59 months), is a specific subset of undernutrition characterized by insufficient height or weight against age-specific growth reference standards[3–5]. The prevalence of stunting, wasting, or underweight in children under five is the proportion of children with a height-for-age, weight-for-height, or weight-for-age *z*-score, respectively, that is more than two standard deviations below the World Health Organization's median growth reference standards for a healthy population[6]. Subnational estimates of CGF report substantial heterogeneity within countries, but are available primarily at the first administrative level (for example, states or provinces)[7]; the uneven geographical distribution of CGF has motivated further calls for assessments that can match the local scale of many public health programmes[8]. Building from our previous work mapping CGF in Africa[9], here we provide the first, to our knowledge, mapped high-spatial-resolution estimates of CGF indicators from 2000 to 2017 across 105 low- and middle-income countries (LMICs), where 99% of affected children live[1], aggregated to policy-relevant first and second (for example, districts or counties) administrative-level units and national levels. Despite remarkable declines over the study period, many LMICs remain far from the ambitious World Health Organization Global Nutrition Targets to reduce stunting by 40% and wasting to less than 5% by 2025. Large disparities in prevalence and progress exist across and within countries; our maps identify high-prevalence areas even within nations otherwise succeeding in reducing overall CGF prevalence. By highlighting where the highest-need populations reside, these geospatial estimates can support policy-makers in planning interventions that are adapted locally and in efficiently directing resources towards reducing CGF and its health implications.

Despite improvements in nearly all LMICs, stunting remained the most widespread and prevalent indicator of CGF throughout the study period. Overall, estimated childhood stunting prevalence across LMICs decreased from 36.9% (95% uncertainty interval, 32.8–41.4%) in 2000 to 26.6% (21.5–32.4%) in 2017. Progress was particularly noticeable in Central America and the Caribbean, Andean South America, North Africa, and East Asia regions, and in some coastal central and western sub-Saharan African (SSA) countries, where most areas with estimated stunting prevalence of at least 50% in 2000 had reduced to 30% or less by 2017 (Fig. 1a, b). By 2017, zones with the highest prevalence of stunting primarily persisted throughout much of the SSA, Central and South Asia, and Oceania regions, where large areas had estimated levels of at least 40%, such as in the first administrative-level units of Nigeria's Jigawa state (60.6% (51.5–69.7%)), Burundi's Karuzi province (60.0% (51.4–67.5%)), India's Uttar Pradesh state (49.0% (48.5–49.5%)), and Laos's Houaphan province (58.3% (50.7–66.8%)) (Extended Data Fig. 1). In 2017, Guatemala (47.0% (40.2–54.6%)), Niger (47.5% (42.2–53.9%)), Burundi (54.2% (46.3–61.2%)), Madagascar (49.8% (43.2–57.2%)), Timor-Leste (49.8% (43.4–56.2%)), and Yemen (45.4% (38.8–51.5%)) had the highest national-level stunting prevalence.

Even within the aforementioned regions where reductions were most evident, local-level estimates revealed communities in which levels still approached those seen in SSA and South Asia; areas in southern Mexico and central Ecuador had estimated stunting prevalence of at least 40%, and areas in western Mongolia reached at least 30%. Wide within-country disparities were apparent in several instances, indicating large areas left behind by the general pace of progress that require attention (Fig. 1a, b). Although most countries successfully reduced

*A list of participants and their affiliations appears in the online version of the paper.

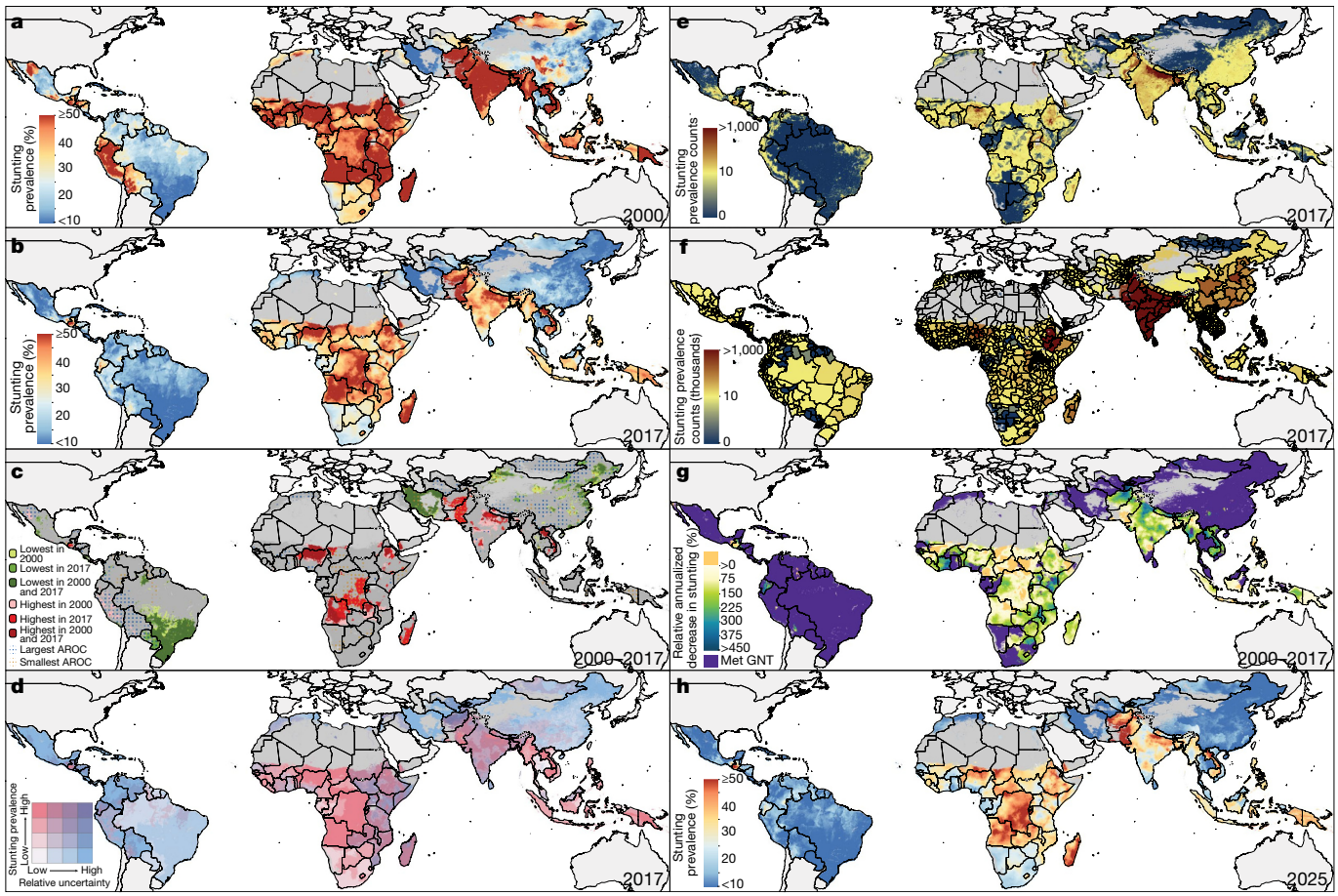


Fig. 1 | Prevalence of stunting in children under five in LMICs (2000–2017) and progress towards 2025. a, **b**, Prevalence of stunting in children under five at the 5 × 5-km resolution in 2000 (**a**) and 2017 (**b**). **c**, Overlapping population-weighted tenth and ninetieth percentiles (lowest and highest) of 5 × 5-km grid cells and AROC in stunting, 2000–2017. **d**, Overlapping population-weighted quartiles of stunting prevalence and relative 95% uncertainty in 2017. **e**, **f**, Number of children under five who were stunted, at the 5 × 5-km (**e**) and first-administrative-unit (**f**) levels. **g**, 2000–2017 annualized decrease in stunting prevalence relative to rates needed during 2017–2025 to meet the WHO GNT. **h**, Grid-cell-level predicted stunting prevalence in 2025. Maps were produced using ArcGIS Desktop 10.6. Interactive visualization tools are available at https://vizhub.healthdata.org/lbd/cgf.

stunting prevalence, subnational inequalities (disparities between second administrative-level units (henceforth 'units')) remained widespread globally—especially evident in Vietnam, Honduras, Nigeria, and India (Extended Data Fig. 2). Among the top quintile of widest disparities, Indonesia experienced a twofold difference in stunting levels in 2017, ranging from 21.0% (16.2–27.0%) in Kota Yogyakarta regency (Yogyakarta province) to 51.5% (40.6–62.3%) in Sumba Barat regency (Nusa Tenggara Timur province). Stunting levels varied fourfold in Nigeria, ranging from 14.7% (9.1–21.0%) in Surulere Local Government Area (Lagos state) to 64.2% (54.2–74.6%) in Gagarawa Local Government Area (Jigawa state) in 2017.

Evaluated from estimates of population-weighted prevalence for areas with the highest and lowest estimated prevalence of stunting (ninetieth and tenth percentiles, respectively), locations in central Chad, Pakistan, and Afghanistan, in northeastern Angola, and throughout the Democratic Republic of the Congo and Madagascar had among the lowest annualized rates of change (AROC), indicating stagnation or increase over the study period (Fig. 1c); in 2017, these countries also had large geographical areas among the most highly prevalent for stunting. By contrast, areas scattered throughout Peru, northwestern Mexico, and eastern Nepal had among the highest stunting levels in 2000, but also the highest rates of decline; by 2017, many of these areas were subsequently no longer in the highest-prevalence decile.

The absolute number of children under five who were stunted was also unequally distributed (Fig. 1e, f), with a large proportion concentrated in a few nations in 2017; overall, 85.1% (84.4–85.7%) of all stunted children under five lived in Africa or Asia. Of the 176.1 million (151.6–203.3 million) children who were stunted in 2017, just over half (50.1% (48.5–52.0%)) lived in only four countries: India (51.5 million (47.7–55.3 million) children; 28.6% (27.1–30.4%) of global stunting), Pakistan (10.7 million (9.3–12.1 million); 6.8% (6.7–6.9%)), Nigeria (11.8 million (10.7–13.0 million); 6.6% (6.4–6.8%)), and China (16.2 million (14.0–18.5 million); 9.0% (9.1–8.9%)). Although China had a low prevalence of national stunting (10.8% (9.1–12.6%)) in 2017, the prevalence was high in India (39.3% (39.1–39.6%)), Pakistan (44.0% (38.4–49.9%)), and Nigeria (38.2% (34.5–42.0%)). Even with moderate levels of stunting (10 to <20%)[10], these highly populous countries would substantially contribute to the global share owing to their population size, and reducing their levels would markedly decrease the number of stunted children.

Childhood wasting was less widespread than stunting (Fig. 2a, b), affecting 8.4% (7.9–9.9%) of children under five in LMICs in 2000, and 6.4% (4.9–7.9%) by 2017. Wasting reached critical levels (at least 15%)[11] nationally in 13 LMICs in 2000 and 7 LMICs in 2017, although only in Mauritania (20.7% (16.5–25.6%)) did all units exceed these levels (Extended Data Fig. 3). Critical wasting prevalence was concentrated in few areas across the globe in 2017, including the peri-Sahelian areas of countries stretching from Mauritania to Sudan, as well as areas in South Sudan, Ethiopia, Kenya, Somalia, Yemen, India, Pakistan, Bhutan, and Indonesia. Most LMICs reduced within-country disparities between

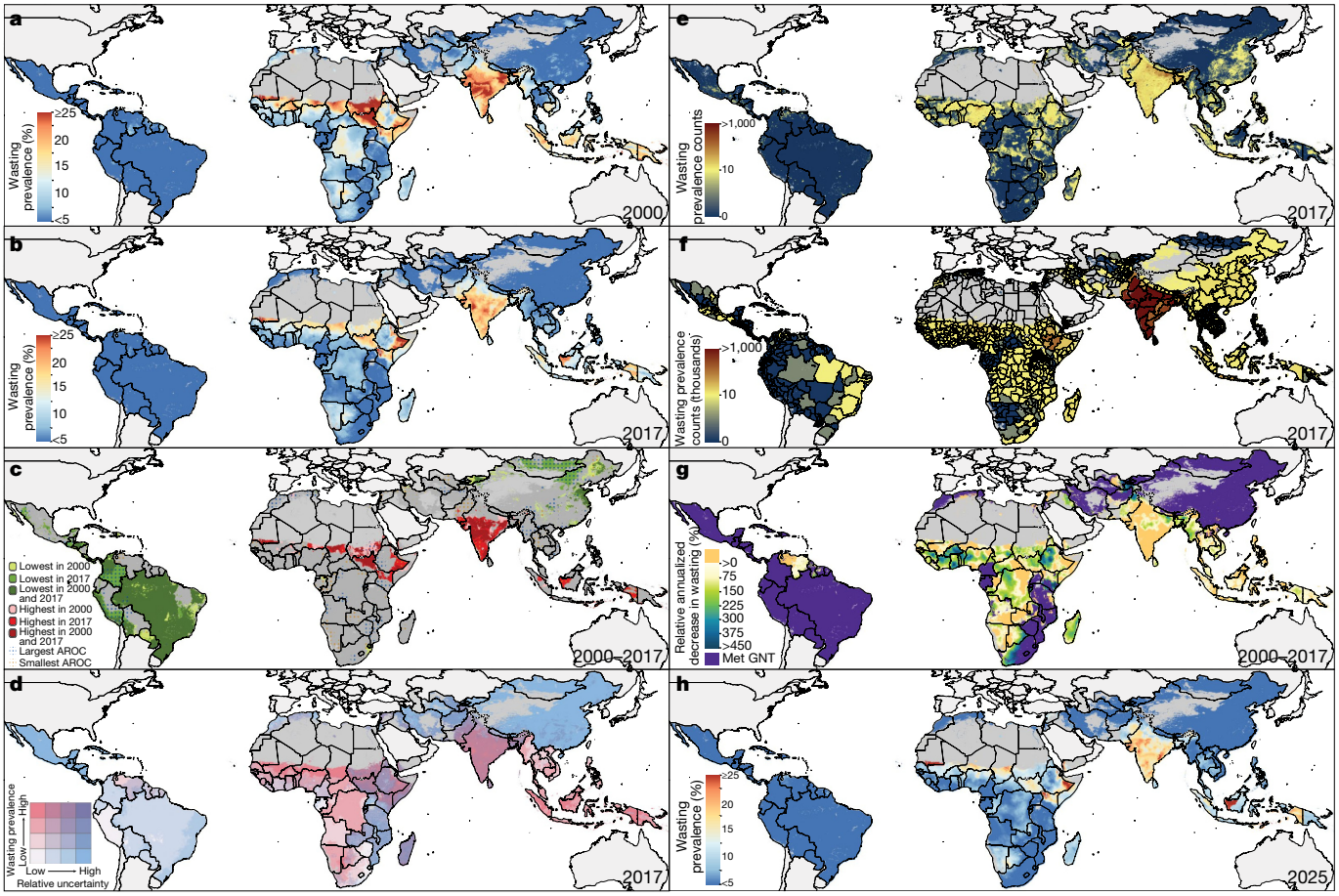


Fig. 2 | Prevalence of wasting in children under five in LMICs (2000–2017) and progress towards 2025. a, **b**, Prevalence of child wasting in children under five at the 5 × 5-km resolution in 2000 (**a**) and 2017 (**b**). **c**, Overlapping population-weighted tenth and ninetieth percentiles (lowest and highest) of 5 × 5-km grid cells and AROC in wasting, 2000–2017. **d**, Overlapping population-weighted quartiles of wasting prevalence and relative 95% uncertainty in 2017. **e**, **f**, Number of children under five affected by wasting, at the 5 × 5-km (**e**) and first-administrative-unit (**f**) levels. **g**, 2000–2017 annualized decrease in wasting prevalence relative to rates needed during 2017–2025 to meet the WHO GNT. **h**, Grid-cell-level predicted wasting prevalence in 2025. Maps were produced using ArcGIS Desktop 10.6. Interactive visualization tools are available at https://vizhub.healthdata.org/lbd/cgf.

their highest- and lowest-prevalence units between 2000 and 2017, most notably in Algeria, Uzbekistan, and Egypt (Extended Data Fig. 4). Even against a backdrop of national-level declines, however, broad within-country disparities in wasting remained in countries such as Indonesia, Ethiopia, Nigeria, and Kenya. An estimated ninefold difference in wasting prevalence occurred among Kenya's units in 2017, ranging from 2.9% (1.6–4.9%) in Tetu constituency (Nyeri county) to 28.3% (20.2–37.3%) in Turkana East constituency (Turkana county); higher-resolution estimates reveal areas with a wasting prevalence of at least 25%. High-prevalence areas in 2000 typically remained within the highest population-weighted decile for wasting in 2017, including the units of Rabkona county (Unity state) in northern South Sudan (27.8% (19.8–37.6%) in 2000; 17.3% (8.8–21.9%) in 2017), the Tanout department (Zinder region) in southern Niger (21.6% (17.3–26.7%) in 2000; 16.5% (11.3–23.3%) in 2017), and Alor regency (Nusa Tenggara Timur province) in southeastern Indonesia (16.4% (9.6–25.8%) in 2000; 20.7% (12.8–30.3%) in 2017) (Fig. 2c).

The absolute number of children affected by wasting was unequal both across and within countries (Fig. 2e, f). Of the 58.3 million (47.6–70.7 million) children affected by wasting in 2017, 57.1% (52.7–61.6%) occurred in four of the most populous countries: India (26.1 million (23.1–29.0 million); 44.7% (41.0–48.6%) of global wasting), Pakistan (3.5 million (2.8–4.3 million); 6.0% (5.8–6.1%)), Bangladesh (1.8 million (1.2–2.4 million); 3.0% (2.6–3.4%)), and Indonesia (2.0 million (1.7–2.3 million); 3.4% (3.3–3.5%)). On the basis of standard thresholds[11], these countries had serious levels of national wasting prevalence (10 to <15%), ranging from 12.2% (9.7–14.9%) in Pakistan to 15.7% (15.5–15.9%) in India, and all but Bangladesh had areas with estimated wasting levels above 20%; increased efforts, especially in densely populated areas with high prevalence and absolute numbers, could immensely reduce global child wasting.

The prevalence of underweight—a composite indicator of stunting and wasting—followed the scattered pattern of high-stunting areas in SSA and spanning Central Asia to Oceania, and the high prevalence belt of wasting along the African Sahel (Extended Data Fig. 5a, b). Affecting 19.8% (17.3–22.7%) of children under five across LMICs in 2000 and 13.0% (10.4–16.0%) in 2017, reductions in underweight prevalence were most notable for countries in Central and South America, southern SSA, North Africa, and Southeast Asia. For example, by 2017, estimated underweight prevalence had decreased to less than or equal to 20% for nearly all areas in Namibia. By contrast, peri-Sahelian countries stretching from Mauritania to Somalia maintained an estimated underweight prevalence of at least 30% in many areas. Large geographical areas across Central and South Asia also maintained high prevalence of underweight during the study period; in particular, India, Pakistan, and Bangladesh sustained estimated prevalence of at least 30% in most locations. Although levels of child underweight had largely reduced since 2000, within-country disparities remained widespread; 71.4% (75 out of 105) of LMICs experienced at least a twofold difference across units in 2017 (Extended Data Fig. 6).

Prospects for reaching 2025 targets

We estimate that broad areas across Central America and the Caribbean, South America, North Africa, and East Asia had high probability (>95%) of having already achieved targets for both stunting and wasting in 2017 (Extended Data Fig. 7). Exceptions to these regional patterns exist; areas with stagnated progress and less than 50% probability of having achieved the World Health Organization's Global Nutrition Targets for 2025 (WHO GNTs) in 2017 were found throughout much of Guatemala and Ecuador for stunting and in southern Venezuela for wasting (Figs. 1g, 2g, Extended Data Fig. 7). Even within countries that had achieved targets, there remain areas with slow progress; locations in central Peru for stunting and southwestern South Africa for wasting had not achieved targets in 2017 (less than 5% probability)—nuances otherwise hidden by aggregated estimates. Owing to stagnation or increases in prevalence, broad areas in SSA and substantial portions across Central Asia, South Asia, and Oceania (for example, in the Democratic Republic of the Congo and Pakistan for stunting; in Yemen and Indonesia for wasting) require reversal of trends or acceleration of declines in order to meet international targets (Figs. 1g, 2g).

Despite predicted improvements in AROC for 2017–2025, many highly affected countries are predicted to have areas that maintain estimated stunting levels of at least 40% or wasting levels of at least 15% in 2025 (Figs. 1h, 2h). Accounting for uncertainty in 2000–2017 AROC estimates, and with 2010 national-level estimates as a baseline for the 40% stunting reduction target, 44.8% (47 out of 105) of LMICs are estimated to nationally meet WHO GNT (>95% probability) for stunting by 2025 (Supplementary Table 13). At finer scales, 17.1% (n = 18) and 7.6% (n = 8) of LMICs will meet the stunting target in all first and second administrative-level units in 2025, respectively (Extended Data Fig. 8a, d, Supplementary Table 13). Similarly, 35.2% (n = 37) of LMICs are estimated to reduce to or maintain less than 5% wasting prevalence by 2025 (>95% probability) based on current trajectories (Supplementary Table 13). Fewer countries were estimated to meet wasting targets in all first administrative-level (16.2% (n = 17)) or second administrative-level (9.5% (n = 10)) units (Extended Data Fig. 8b, e, Supplementary Table 13). Only 26.7% (n = 28) of LMICs will meet national-level targets for both stunting and wasting by 2025, and only 4.8% (n = 5) will achieve both targets in all units (Supplementary Table 13).

Discussion

Although commendable declines in CGF have occurred globally, this progress measured at a coarse scale conceals subnational and local underachievement and variation in achieving the WHO GNTs. Supporting conclusions in the Global Nutrition Report[12], our results show that most LMICs will not reach WHO GNTs nationally, and even fewer will meet targets across subnational units. Our mapped results show broad heterogeneity across areas, and reveal hotspots of persistent CGF even within well-performing regions and countries, where increased and targeted efforts are needed. In 2017, one in four children under five across LMICs still suffered at least one dimension of CGF, and the largest numbers of affected children were often in specific within-country locations. Although the national prevalence of CGF was generally lower in Central America and the Caribbean, South American, and East Asian countries, there are communities in these regions in which levels of CGF remain as high as those in SSA and South Asia. Regardless of overall declines, many subnational areas across LMICs maintained high levels of CGF and require substantial acceleration of progress or reversal of increasing trends to meet nutrition targets and leave no populations behind.

To our knowledge, this study is the first to estimate CGF comprehensively across LMICs at a fine geospatial scale, providing a precision public health tool to support efficient targeting of local-level interventions to vulnerable populations. Although densely populated areas may have relatively low prevalence of CGF, the absolute number of affected children may still be high; thus, both relative and absolute estimates are important to determine where additional attention is needed. To achieve international goals, more concerted efforts are needed in areas with decreasing or stagnating trends, without diminishing support in areas that demonstrate progress nor contributing to increases in obesity. In future work, we plan to determine how to stratify our estimates of CGF by sex and age, assess the double burden of child undernutrition and overweight, analyse important maternal indicators that affect child nutritional status outcomes (such as anaemia), and continue to monitor progress towards the 2025 WHO GNTs. These mapped estimates enable decision-makers to visualize and compare subnational CGF and nutritional inequalities, and identify populations most in need of interventions[13].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1878-8.

1. Dicker, D. et al. Global, regional, and national age-sex-specific mortality and life expectancy, 1950–2017: a systematic analysis for the Global Burden of Disease Study 2017. *Lancet* **392**, 1684–1735 (2018).
2. Victora, C. G. et al. Maternal and child undernutrition: consequences for adult health and human capital. *Lancet* **371**, 340–357 (2008).
3. WHO & UNICEF. *WHO Child Growth Standards and the Identification of Severe Acute Malnutrition in Infants and Children: A Joint Statement* https://www.who.int/nutrition/publications/severemalnutrition/9789241598163/en/ (2009).
4. Wang, Y. & Chen, H.-J. In *Handbook of Anthropometry* (ed. Preedy, V. R.) **2**, 29–48 (Springer New York, 2012).
5. Waterlow, J. C. et al. The presentation and use of height and weight data for comparing the nutritional status of groups of children under the age of 10 years. *Bull. World Health Organ.* **55**, 489–498 (1977).
6. WHO Multicentre Growth Reference Study Group. WHO Child Growth Standards based on length/height, weight and age. *Acta Paediatr.* **450**, 76–85 (2006).
7. ICF & USAID. *The DHS Program: Demographic and Health Surveys* https://dhsprogram.com/publications/Publication-Search.cfm?shareurl=yes&topic1=15&pubTypeSelected=pubtype_5 (accessed 13 September 2018).
8. Reich, B. J. & Haran, M. Precision maps for public health. *Nature* **555**, 32–33 (2018).
9. Osgood-Zimmerman, A. et al. Mapping child growth failure in Africa between 2000 and 2015. *Nature* **555**, 41–47 (2018).
10. de Onis, M. et al. Prevalence thresholds for wasting, overweight and stunting in children under 5 years. *Public Health Nutr.* **22**, 1–5 (2018).
11. WHO. *Nutrition Landscape Information System (NLIS) Country Profile Indicators Interpretation Guide* https://www.who.int/nutrition/nlis_interpretationguide_isbn9789241599955/en/ (2010).
12. Development Initiatives. *The 2018 Global Nutrition Report: Shining a Light to Spur Action on Nutrition* https://globalnutritionreport.org/reports/global-nutrition-report-2018/ (2018).
13. Annan, K. Data can help to end malnutrition across Africa. *Nature* **555**, 7 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2020

Local Burden of Disease Child Growth Failure Collaborators

Damaris K. Kinyoki[1,2], Aaron E. Osgood-Zimmerman[1], Brandon V. Pickering[1], Lauren E. Schaeffer[1], Laurie B. Marczak[1], Alice Lazzar-Atwood[1], Michael L. Collison[1], Nathaniel J. Henry[1], Zegeye Abebe[3], Abdu A. Adamu[4,5], Victor Adekanmbi[6], Keivan Ahmadi[7], Olufemi Ajumobi[8,9], Ayman Al-Eyadhy[10], Rajaa M. Al-Raddadi[11], Fares Alahdab[12], Mehran Alijanzadeh[13], Vahid Alipour[14,15], Khalid Altirkawi[16], Saeed Amini[17], Catalina Liliana Andrei[18], Carl Abelardo T. Antonio[19,20], Jalal Arabloo[15], Olatunde Aremu[21], Mehran Asadi-Aliabadi[22], Suleman Atique[23], Marcel Ausloos[24,25], Marco Avila[26], Ashish Awasthi[27,28], Beatriz Paulina Ayala Quintanilla[29,30], Samad Azari[15], Alaa Badawi[31,32], Till Winfried Bärnighausen[33,34], Quique Bassat[35,36], Kaleab Baye[37], Neeraj Bedi[38,39], Bayu Begashaw Bekele[40,41], Michelle L. Bell[42], Natalia V. Bhattacharjee[1], Krittika Bhattacharyya[43,44], Suraj Bhattarai[45], Zulfiqar A. Bhutta[46,47], Belete Biadgo[48], Boris Bikbov[49], Andrey Nikolaevich Briko[50], Gabrielle Britton[51], Roy Burstein[1], Zahid A. Butt[52,53], Josip Car[54,55], Carlos A. Castañeda-Orjuela[56,57], Franz Castro[58], Ester Cerin[59,60], Michael G. Chipeta[61], Dinh-Toi Chu[62], Michael A. Cork[1], Elizabeth A. Cromwell[1,2], Lucía Cuevas-Nasu[26], Lalit Dandona[1,28], Rakhi Dandona[1,28], Farah Daoud[1], Rajat Das Gupta[63,64], Nicole Davis Weaver[1], Diego De Leo[65], Jan-Walter De Neve[33], Kebede Deribe[66,67], Beruk Berhanu Desalegn[68], Aniruddha Deshpande[1], Melaku Desta[69,70], Daniel Diaz[70,71], Mesfin Tadese Dinberu[72], David Teye Doku[73,74], Manisha Dubey[75], Andre R. Durães[76,77], Laura Dwyer-Lindgren[1,2], Lucas Earl[1], Andem Effiong[78], Maysaa El Sayed Zaki[79], Maha El Tantawi[80], Ziad El-Khatib[81,82], Babak Eshrati[83,84], Mohammad Fareed[85], Andre Faro[86], Seyed-Mohammad Fereshtehnejad[87,88], Irina Filip[89,90], Florian Fischer[91], Nataliya A. Foigt[92], Morenike Oluwatoyin Folayan[93], Takeshi Fukumoto[94,95], Tsegaye Tewelde Gebrehiwot[96], Kebede Embaye Gezae[97], Alireza Ghajar[98,99], Paramjit Singh Gill[100], Philimon N. Gona[101], Sameer Vali Gopalani[102,103], Ayman Grada[104], Yuming Guo[105,106], Arvin Haj-Mirzaian[107,108], Arya Haj-Mirzaian[107,109], Jason B. Hall[1], Samer Hamidi[110], Andualem Henok[41], Bernardo Hernández Prado[1,2], Mario Herrero[111], Claudiu Herteliu[112], Chi Linh Hoang[113], Michael K. Hole[114], Naznin Hossain[115,116], Mehdi Hosseinzadeh[117,118], Guoqing Hu[119], Sheikh Mohammed Shariful Islam[120,121], Mihajlo Jakovljevic[122], Ravi Prakash Jha[123], Jost B. Jonas[124,125], Jacek Jerzy Jozwiak[126], Amaha Kahsay[127], Tanuj Kanchan[128], Manoochehr Karami[129], Amir Kasaeian[130,131], Yousef Saleh Khader[132], Ejaz Ahmad Khan[133], Mona M. Khater[134], Yun Jin Kim[135], Ruth W. Kimokoti[136], Adnan Kisa[137], Sonali Kochhar[138,139], Soewarta Kosen[140], Ai Koyanagi[36,141], Kewal Krishan[142], Barthelemy Kuate Defo[143,144], G. Anil Kumar[28], Manasi Kumar[145,146], Sheetal D. Lad[147], Faris Hasan Lami[148], Paul H. Lee[149], Aubrey J. Levine[1], Shanshan Li[105], Shai Linn[150], Rakesh Lodha[151], Hassan Magdy Abd El Razek[152], Muhammed Magdy Abd El Razek[153], Marek Majdan[154], Azeem Majeed[155], Reza Malekzadeh[156,157], Deborah Carvalho Malta[158], Abdullah A. Mamun[159], Mohammad Ali Mansournia[160], Francisco Rogerlândio Martins-Melo[161], Anthony Masaka[162], Benjamin Ballard Massenburg[163], Benjamin K. Mayala[1], Fabiola Mejia-Rodriguez[164], Mulugeta Melku[40], Walter Mendoza[165], George A. Mensah[166,167], Tomasz Miazgowski[168], Ted R. Miller[169,170], G. K. Mini[171,172], Erkin M. Mirrakhimov[173,174], Babak Moazen[33,175], Aso Mohammad Darwesh[176], Shafiu Mohammed[33,177], Farnam Mohebi[178], Ali H. Mokdad[1,2], Yoshan Moodley[179], Ghobad Moradi[180,181], Maziar Moradi-Lakeh[22], Paula Moraga[182], Shane Douglas Morrison[183], Jonathan F. Mosser[1], Seyyed Meysam Mousavi[184,185], Ulrich Otto Mueller[186,187], Christopher J. L. Murray[1,2], Ghulam Mustafa[188,189], Mehdi Naderi[190], Mohsen Naghavi[1,2], Farid Najafi[191], Vinay Nangia[192], Duduzile Edith Ndwandwe[5], Ionut Negoi[193], Josephine W. Ngunjiri[194], Huong Lan Thi Nguyen[195], Long Hoang Nguyen[113], Son Hoang Nguyen[113], Jing Nie[196], Chukwudi A. Nnaji[5,197], Jean Jacques Noubiap[167], Malihe Nourollahpour Shiadeh[198], Peter S. Nyasulu[199], Felix Akpojene Ogbo[200], Andrew T. Olagunju[201,202], Bolajoko Olubukunola Olusanya[203], Jacob Olusegun Olusanya[203], Eduardo Ortiz-Panozo[204,205], Stanislav S. Otstavnov[206,207], Mahesh P. A.[208], Adrian Pana[112,209], Anamika Pandey[28], Sanghamitra Pati[210], Snehal T. Patil[211], George C. Patton[212,213], Norberto Perico[214], David M. Pigott[1,2], Meghdad Pirsaheb[215], Ellen G. Piwoz[216], Maarten J. Postma[217,218], Akram Pourshams[156], Swayam Prakash[219], Hedley Quintana[58], Amir Radfar[220,221], Alireza Rafiei[222,223], Vafa Rahimi-Movaghar[224], Rajesh Kumar Rai[225,226], Fatemeh Rajati[215], David Laith Rawaf[227,228], Salman Rawaf[229,230], Rahul Rawat[216], Giuseppe Remuzzi[214], Andre M. N. Renzaho[231,232], Carlos Rios-González[233,234], Leonardo Roever[235], Jennifer M. Ross[1,138], Ali Rostami[236], Nafis Sadat[1], Yahya Safari[215], Mahdi Safdarian[224,237], Amirhossein Sahebkar[238,239], Nasir Salam[240], Payman Salamati[224], Yahya Salimi[191,241], Hamideh Salimzadeh[156], Abdallah M. Samy[242], Benn Sartorius[2,243], Brijesh Sathian[244,245], Megan F. Schipp[1], David C. Schwebel[246], Anbissa Muleta Senbeta[247], Sadaf G. Sepanlou[156,157], Masood Ali Shaikh[248], Teresa Shamah Levy[26], Mohammadbagher Shamsi[249], Kiomars Sharafi[215], Rajesh Sharma[250], Aziz Sheikh[251,252], Apurba Shil[253], Diego Augusto Santos Silva[254], Jasvinder A. Singh[255,256], Dhirendra Narain Sinha[257,258], Moslem Soofi[241], Agus Sudaryanto[259,260], Mu'awiyyah Babale Sufiyan[261], Rafael Tabarés-Seisdedos[262,263], Birkneh Tilahun Tadesse[264,265,267], Mohamad-Hani Temsah[266,267], Abdullah Sulieman Terkawi[268,269], Zemenu Tadesse Tessema[270], Andrew L. Thorne-Lyman[271], Marcos Roberto Tovani-Palone[272], Bach Xuan Tran[273], Khanh Bao Tran[274,275], Irfan Ullah[276,277], Olalekan A. Uthman[278], Masoud Vaezghasemi[279], Afsane Vaezi[280], Pascual R. Valdez[281,282], John Vanderheide[1], Yousef Veisani[283], Francesco S. Violante[284,285], Vasily Vlassov[286], Giang Thu Vu[113], Linh Gia Vu[113], Yasir Waheed[287], Judd L. Walson[138], Yafeng Wang[288], Yuan-Pang Wang[289], Elizabeth N. Wangia[290], Andrea Werdecker[186,187], Gelin Xu[291], Tomohide Yamada[292], Engida Yisma[293], Naohiro Yonemoto[294], Mustafa Z. Younis[295,296], Mahmoud Yousefifard[297], Chuanhua Yu[288,298], Sojib Bin Zaman[299,300], Mohammad Zamani[301], Yunquan Zhang[302,303], Nicholas J. Kassebaum[1,304,305] & Simon I. Hay[1,2,305]*

[1]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA. [2]Department of Health Metrics Sciences, School of Medicine, University of Washington, Seattle, WA, USA. [3]Human Nutrition Department, University of Gondar, Gondar, Ethiopia. [4]Department of Global Health, Stellenbosch University, Cape Town, South Africa. [5]Cochrane South Africa, South African Medical Research Council, Cape Town, South Africa. [6]School of Medicine, Cardiff University, Cardiff, UK. [7]Lincoln Medical School, Universities of Nottingham & Lincoln, Lincoln, UK. [8]School of Community Health Sciences, University of Nevada, Reno, NV, USA. [9]National Malaria Elimination Program, Federal Ministry of Health, Abuja, Nigeria. [10]Pediatric Intensive Care Unit, King Saud University, Riyadh, Saudi Arabia. [11]Department of Family and Community Medicine, King Abdulaziz University, Jeddah, Saudi Arabia. [12]Evidence Based Practice Center, Mayo Clinic Foundation for Medical Education and Research, Rochester, MN, USA. [13]Qazvin University of Medical Sciences, Qazvin, Iran. [14]Health Economics Department, Iran University of Medical Sciences, Tehran, Iran. [15]Health Management and Economics Research Center, Iran University of Medical Sciences, Tehran, Iran. [16]King Saud University, Riyadh, Saudi Arabia. [17]Health Services Management Department, Arak University of Medical Sciences, Arak, Iran. [18]Carol Davila University of Medicine & Pharmacy, Bucharest, Romania. [19]Department of Health Policy & Administration, University of the Philippines Manila, Manila, The Philippines. [20]Department of Applied Social Sciences, Hong Kong Polytechnic University, Hong Kong, China. [21]School of Health Sciences, Birmingham City University, Birmingham, UK. [22]Preventive Medicine and Public Health Research Center, Iran University of Medical Sciences, Tehran, Iran. [23]Department of Health Informatics, University of Ha'il, Ha'il, Saudi Arabia. [24]School of Business, University of Leicester, Leicester, UK. [25]Department of Statistics and Econometrics, Bucharest University of Economic Studies, Bucharest, Romania. [26]Center for Research in Evaluation and Surveys, National Public Health Institute, Cuernavaca, Mexico. [27]Indian Institute of Public Health, Gandhinagar, India. [28]Public Health Foundation of India, Gurugram, India. [29]The Judith Lumley Centre, La Trobe University, Melbourne, Victoria, Australia. [30]General Office for Research and Technological Transfer, Peruvian National Institute of Health, Lima, Peru. [31]Public Health Risk Sciences Division, Public Health Agency of Canada, Toronto, Ontario, Canada. [32]Department of Nutritional Sciences, University of Toronto, Toronto, Ontario, Canada. [33]Heidelberg Institute of Global Health (HIGH), Heidelberg University, Heidelberg, Germany. [34]T. H. Chan School of Public Health, Harvard University, Boston, MA, USA. [35]Barcelona Institute for Global Health, University of Barcelona, Barcelona, Spain. [36]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. [37]Center for Food Science and Nutrition, Addis Ababa University, Addis Ababa, Ethiopia. [38]Department of Community Medicine, Gandhi Medical College Bhopal, Bhopal, India. [39]Jazan University, Jazan, Saudi Arabia. [40]Institute of Public Health, University of Gondar, Gondar, Ethiopia. [41]Public Health Department, Mizan-Tepi University, Teppi, Ethiopia. [42]School of Forestry and Environmental Studies, Yale University, New Haven, CT, USA. [43]Department of Statistical and Computational Genomics, National Institute of Biomedical Genomics, Kalyani, India. [44]Department of Statistics, University of Calcutta, Kolkata, India. [45]Department of Global Health, Global Institute for Interdisciplinary Studies, Kathmandu, Nepal. [46]Centre for Global Child Health, University of Toronto, Toronto, Ontario, Canada. [47]Centre of Excellence in Women and Child Health, Aga Khan University, Karachi, Pakistan. [48]Department of Clinical Chemistry, University of Gondar, Gondar, Ethiopia. [49]Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Ranica, Italy. [50]Biomedical Technologies, Bauman Moscow State Technical University, Moscow, Russia. [51]Center for Neuroscience, Instituto de Investigaciones Científicas y Servicios de Alta Tecnología (INDICASAT AIP), Panama, Panama. [52]School of Public Health and Health Systems, University of Waterloo, Waterloo, Ontario, Canada. [53]Al Shifa School of Public Health, Al Shifa Trust Eye Hospital, Rawalpindi, Pakistan. [54]Centre for Population Health Sciences, Nanyang Technological University, Singapore, Singapore. [55]Global Health Unit, Imperial College London, London, UK. [56]Colombian National Health Observatory, National Institute of Health, Bogota, Colombia. [57]Epidemiology and Public Health Evaluation Group, National University of Colombia, Bogota, Colombia. [58]Gorgas Memorial Institute for Health Studies, Panama, Panama. [59]Mary Mackillop Institute for Health Research, Australian Catholic University, Melbourne, Victoria, Australia. [60]School of Public Health, University of Hong Kong, Hong Kong, China. [61]Big Data Institute, University of Oxford, Oxford, UK. [62]Faculty of Biology, Hanoi National University of Education, Hanoi, Vietnam. [63]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, USA. [64]James P. Grant School of Public Health, BRAC University, Dhaka, Bangladesh. [65]Australian Institute for Suicide Research and Prevention, Griffith University, Mount Gravatt, Queensland, Australia. [66]School of Public Health, Addis Ababa University, Addis Ababa, Ethiopia. [67]Department of Global Health and Infection, Brighton and Sussex Medical School, Brighton, UK. [68]School of Nutrition, Food Science and Technology, Hawassa University, Hawassa, Ethiopia. [69]Department of Midwifery, Debre Markos University, Debre Markos, Ethiopia. [70]Faculty of Veterinary Medicine and Zootechnics, Autonomous University of Sinaloa, Culiacan Rosales, Mexico. [71]Center of Complexity Sciences, National Autonomous University of Mexico, Mexico City, Mexico. [72]Department of Midwifery, Debre Berhan University, Debre Berhan, Ethiopia. [73]Department of Population and Health, University of Cape Coast, Cape Coast, Ghana. [74]Faculty of Social Sciences, Health Sciences, University of Tampere, Tampere, Finland. [75]World Food Programme, New Delhi, India. [76]Medical Board, Roberto Santos General Hospital, Salvador, Brazil. [77]Department of Internal Medicine, Bahia School of Medicine and Public Health, Salvador, Brazil. [78]Clinical Epidemiology and Biostatistics, University of Newcastle, Newcastle, New South Wales, Australia. [79]Department of Clinical Pathology, Mansoura University, Mansoura, Egypt. [80]Pediatric Dentistry and Dental Public Health, Alexandria University, Alexandria, Egypt. [81]Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden. [82]World Health Programme, Université du Québec en Abitibi-Témiscamingue, Rouyn-Noranda, Quebec, Canada. [83]School of Public Health, Arak University of Medical Sciences, Arak, Iran. [84]Center of Communicable Disease Control, Ministry of Health and Medical

Education, Tehran, Iran. [85]College of Medicine, Imam Muhammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. [86]Department of Psychology, Federal University of Sergipe, Sao Cristovao, Brazil. [87]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden. [88]Division of Neurology, University of Ottawa, Ottawa, Ontario, Canada. [89]Psychiatry Department, Kaiser Permanente, Fontana, CA, USA. [90]Department of Health Sciences, A. T. Still University, Mesa, AZ, USA. [91]Department of Population Health Medicine and Health Services Research, Bielefeld University, Bielefeld, Germany. [92]Laboratory of Population Aging, Institute of Gerontology, National Academy of Medical Sciences of Ukraine, Kyiv, Ukraine. [93]Department of Child Dental Health, Obafemi Awolowo University, Ile-Ife, Nigeria. [94]Gene Expression & Regulation Program, The Wistar Institute, Philadelphia, PA, USA. [95]Department of Dermatology, Kobe University, Kobe, Japan. [96]Department of Epidemiology, Jimma University, Jimma, Ethiopia. [97]Department of Biostatistics, Mekelle University, Mekelle, Ethiopia. [98]Endocrinology and Metabolism Research Center (EMRC), Tehran University of Medical Sciences, Tehran, Iran. [99]Department of Medicine, Massachusetts General Hospital, Boston, MA, USA. [100]Unit of Academic Primary Care, University of Warwick, Coventry, UK. [101]Nursing and Health Sciences Department, University of Massachusetts Boston, Boston, MA, USA. [102]Department of Biostatistics and Epidemiology, University of Oklahoma, Oklahoma City, OK, USA. [103]Department of Health and Social Affairs, Government of the Federated States of Micronesia, Palikir, Federated States of Micronesia. [104]Department of Dermatology, Boston University, Boston, MA, USA. [105]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia. [106]Department of Epidemiology and Biostatistics, College of Public Health, Zhengzhou University, Zhengzhou, China. [107]Department of Pharmacology, Tehran University of Medical Sciences, Tehran, Iran. [108]Obesity Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [109]Department of Radiology, Johns Hopkins University, Baltimore, MD, USA. [110]School of Health and Environmental Studies, Hamdan Bin Mohammed Smart University, Dubai, United Arab Emirates. [111]Agriculture and Food, Commonwealth Scientific and Industrial Research Organisation, St Lucia, Queensland, Australia. [112]Department of Statistics and Econometrics, Bucharest University of Economic Studies, Bucharest, Romania. [113]Center of Excellence in Behavioral Medicine, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam. [114]Department of Pediatrics, Dell Medical School, University of Texas Austin, Austin, TX, USA. [115]Department of Pharmacology and Therapeutics, Dhaka Medical College, Dhaka, Bangladesh. [116]Department of Pharmacology, Bangladesh Industrial Gases Limited, Tangail, Bangladesh. [117]Department of Computer Engineering, Islamic Azad Univeristy, Tehran, Iran. [118]Computer Science Department, University of Human Development, Sulaimaniyah, Iraq. [119]Department of Epidemiology and Health Statistics, Central South University, Changsha, China. [120]Institute for Physical Activity and Nutrition, Deakin University, Burwood, Victoria, Australia. [121]Sydney Medical School, University of Sydney, Sydney, New South Wales, Australia. [122]Department of Health Care and Public Health, Sechenov First Moscow State Medical University, Moscow, Russia. [123]Department of Community Medicine, Banaras Hindu University, Varanasi, India. [124]Department of Ophthalmology, Heidelberg University, Heidelberg, Germany. [125]Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Beijing, China. [126]Department of Family Medicine and Public Health, University of Opole, Opole, Poland. [127]Department of Nutrition and Dietetics, Mekelle University, Mekelle, Ethiopia. [128]Department of Forensic Medicine and Toxicology, All India Institute of Medical Sciences, Jodhpur, India. [129]Department of Epidemiology, Hamadan University of Medical Sciences, Hamadan, Iran. [130]Pars Advanced and Minimally Invasive Medical Manners Research Center, Iran University of Medical Sciences Tehran, Tehran, Iran. [131]Hematology-Oncology and Stem Cell Transplantation Research Center, Tehran University of Medical Sciences, Tehran, Iran. [132]Department of Public Health, Jordan University of Science and Technology, Irbid, Jordan. [133]Epidemiology and Biostatistics Department, Health Services Academy, Islamabad, Pakistan. [134]Department of Medical Parasitology, Cairo University, Cairo, Egypt. [135]School of Medicine, Xiamen University Malaysia, Sepang, Malaysia. [136]Department of Nutrition, Simmons University, Boston, MA, USA. [137]School of Health Sciences, Kristiania University College, Oslo, Norway. [138]Department of Global Health, University of Washington, Seattle, WA, USA. [139]Department of Public Health, Erasmus University Medical Center, Rotterdam, The Netherlands. [140]Independent Consultant, Jakarta, Indonesia. [141]CIBERSAM, San Juan de Dios Sanitary Park, Sant Boi De Llobregat, Spain. [142]Department of Anthropology, Panjab University, Chandigarh, India. [143]Department of Social and Preventive Medicine, University of Montreal, Montreal, Quebec, Canada. [144]Department of Demography, University of Montreal, Montreal, Quebec, Canada. [145]Department of Psychiatry, University of Nairobi, Nairobi, Kenya. [146]Division of Psychology and Language Sciences, University College London, London, UK. [147]Department of Pediatrics, Post Graduate Institute of Medical Education and Research, Chandigarh, India. [148]Department of Community and Family Medicine, University of Baghdad, Baghdad, Iraq. [149]School of Nursing, Hong Kong Polytechnic University, Hong Kong, China. [150]School of Public Health, University of Haifa, Haifa, Israel. [151]Department of Paediatrics, All India Institute of Medical Sciences, Jodhpur, India. [152]Radiology Department, Mansoura Faculty of Medicine, Mansoura, Egypt. [153]Ophthalmology Department, Aswan Faculty of Medicine, Aswan, Egypt. [154]Department of Public Health, Trnava University, Trnava, Slovakia. [155]Department of Primary Care and Public Health, Imperial College London, London, UK. [156]Digestive Diseases Research Institute, Tehran University of Medical Sciences, Tehran, Iran. [157]Non-communicable Diseases Research Center, Shiraz University of Medical Sciences, Shiraz, Iran. [158]Department of Maternal and Child Nursing and Public Health, Federal University of Minas Gerais, Belo Horizonte, Brazil. [159]Institute for Social Science Research, The University of Queensland, Brisbane, Queensland, Australia. [160]Department of Epidemiology and Biostatistics, Tehran University of Medical Sciences, Tehran, Iran. [161]Campus Caucaia, Federal Institute of Education, Science and Technology of Ceará, Caucaia, Brazil. [162]Public Health Department, Botho University-Botswana, Gaborone, Botswana. [163]Division of Plastic Surgery, University of Washington, Seattle, WA, USA. [164]Research in Nutrition and Health, National Institute of Public Health, Cuernavaca, Mexico. [165]Peru Country Office, United Nations Population Fund (UNFPA), Lima, Peru. [166]Center for Translation Research and Implementation Science, National Institutes of Health, Bethesda, MD, USA. [167]Department of Medicine, University of Cape Town, Cape Town, South Africa. [168]Department of Propedeutics of Internal Diseases & Arterial Hypertension, Pomeranian Medical University, Szczecin, Poland. [169]Pacific Institute for Research & Evaluation, Calverton, MD, USA. [170]School of Public Health, Curtin University, Perth, Western Australia, Australia. [171]Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India. [172]Global Institute of Public Health (GIPH), Ananthapuri Hospitals and Research Centre, Trivandrum, India. [173]Faculty of Internal Medicine, Kyrgyz State Medical Academy, Bishkek, Kyrgyzstan. [174]Department of Atherosclerosis and Coronary Heart Disease, National Center of Cardiology and Internal Disease, Bishkek, Kyrgyzstan. [175]Institute of Addiction Research (ISFF), Frankfurt University of Applied Sciences, Frankfurt, Germany. [176]Department of Information Technology, University of Human Development, Sulaymaniyah, Iraq. [177]Health Systems and Policy Research Unit, Ahmadu Bello University, Zaria, Nigeria. [178]Non-communicable Diseases Research Center, Tehran University of Medical Sciences, Tehran, Iran. [179]Department of Public Health Medicine, University of Kwazulu-Natal, Durban, South Africa. [180]Department of Epidemiology and Biostatistics, Kurdistan University of Medical Sciences, Sanandaj, Iran. [181]Social Determinants of Health Research Center, Kurdistan University of Medical Sciences, Sanandaj, Iran. [182]Department of Mathematical Sciences, University of Bath, Bath, UK. [183]Department of Surgery, University of Washington, Seattle, WA, USA. [184]Department of Health Management and Economics, Tehran University of Medical Sciences, Tehran, Iran. [185]Health Management Research Center, Baqiyatallah Univeristy of Medical Sciences, Tehran, Iran. [186]Federal Institute for Population Research, Wiesbaden, Germany. [187]Center for Population and Health, Wiesbaden, Germany. [188]Department of Pediatric Medicine, Nishtar Medical University, Multan, Pakistan. [189]Department of Pediatrics & Pediatric Pulmonology, Institute of Mother & Child Care, Multan, Pakistan. [190]Clinical Research Development Centre, Kermanshah University of Medical Sciences, Kermanshah, Iran. [191]Department of Epidemiology & Biostatistics, Kermanshah University of Medical Sciences, Kermanshah, Iran. [192]Suraj Eye Institute, Nagpur, India. [193]General Surgery, Carol Davila University of Medicine & Pharmacy, Bucharest, Romania. [194]Department of Biological Sciences, University of Embu, Embu, Kenya. [195]Institute for Global Health Innovations, Duy Tan University, Hanoi, Vietnam. [196]Department of Sociology & Institute for Empirical Social Science Research, Xi'an Jiaotong University, Xi'an, China. [197]School of Public Health and Family Medicine, University of Cape Town, Cape Town, South Africa. [198]Mazandaran University of Medical Sciences, Sari, Iran. [199]Faculty of Medicine & Health Sciences, Stellenbosch University, Cape Town, South Africa. [200]UCIBIO, University of Porto, Porto, Portugal. [201]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, Canada. [202]Department of Psychiatry, University of Lagos, Lagos, Nigeria. [203]Centre for Healthy Start Initiative, Lagos, Nigeria. [204]Center for Population Health Research, National Institute of Public Health, Cuernavaca, Mexico. [205]School of Health and Welfare, Jönköping University, Jönköping, Sweden. [206]Laboratory of Public Health Indicators Analysis and Health Digitalization, Moscow Institute of Physics and Technology, Dolgoprudny, Russia. [207]Department of Project Management, National Research University Higher School of Economics, Moscow, Russia. [208]Department of Respiratory Medicine, Jagadguru Sri Shivarathreeswara Academy of Health Education and Research, Mysore, India. [209]Center for Health Outcomes & Evaluation, Bucharest, Romania. [210]Regional Medical Research Centre, Indian Council of Medical Research, Bhubaneswar, India. [211]Krishna Institute of Medical Sciences, Deemed University, Karad, India. [212]Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia. [213]Population Health, Murdoch Children's Research Institute, Melbourne, Victoria, Australia. [214]Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Bergamo, Italy. [215]Research Center for Environmental Determinants of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran. [216]Bill & Melinda Gates Foundation, Seattle, WA, USA. [217]Department of Economics and Business, University of Groningen, Groningen, The Netherlands. [218]University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [219]Department of Nephrology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, India. [220]College of Graduate Health Sciences, A. T. Still University, Mesa, AZ, USA. [221]College of Medicine, University of Central Florida, Orlando, FL, USA. [222]Molecular and Cell Biology Research Center, Mazandaran University of Medical Sciences, Sari, Iran. [223]Department of Immunology, Mazandaran University of Medical Sciences, Sari, Iran. [224]Sina Trauma and Surgery Research Center, Tehran University of Medical Sciences, Tehran, Iran. [225]Society for Health and Demographic Surveillance, Suri, India. [226]Department of Economics, University of Göttingen, Göttingen, Germany. [227]WHO Collaborating Centre for Public Health Education and Training, Imperial College London, London, UK. [228]University College London Hospitals, London, UK. [229]Academic Public Health, Public Health England, London, UK. [230]Department of Primary Care and Public Health, School of Public Health, Imperial College London, London, UK. [231]School of Social Sciences and Psychology, Western Sydney University, Penrith, New South Wales, Australia. [232]Translational Health Research Institute, Western Sydney University, Penrith, New South Wales, Australia. [233]Research Directorate, Nihon Gakko University, Fernando De La Mora, Paraguay. [234]Research Direction, Universidad Nacional de Caaguazú, Coronel Oviedo, Paraguay. [235]Department of Clinical Research, Federal University of Uberlândia, Uberlândia, Brazil. [236]Infectious Diseases and Tropical Medicine Research Center, Babol University of Medical Sciences, Babol, Iran. [237]Department of Neuroscience, Iran University of Medical Sciences, Tehran, Iran. [238]Neurogenic Inflammation Research Center, Mashhad University of Medical Sciences, Mashhad, Iran. [239]Halal Research Center of IRI, FDA,

Tehran, Iran. [240]Department of Pathology, Al-Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. [241]Social Development & Health Promotion Research Center, Kermanshah University of Medical Sciences, Kermanshah, Iran. [242]Department of Entomology, Ain Shams University, Cairo, Egypt. [243]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK. [244]Surgery Department, Hamad Medical Corporation, Doha, Qatar. [245]Faculty of Health & Social Sciences, Bournemouth University, Bournemouth, UK. [246]Department of Psychology, University of Alabama at Birmingham, Birmingham, AL, USA. [247]Department of Food Science and Nutrition, Jigjiga University, Jigjiga, Ethiopia. [248]Independent Consultant, Karachi, Pakistan. [249]Department of Sports Medicine & Rehabilitation, Kermanshah University of Medical Sciences, Kermanshah, Iran. [250]University School of Management and Entrepreneurship, Delhi Technological University, New Delhi, India. [251]Division of General Internal Medicine, Harvard University, Boston, MA, USA. [252]Centre for Medical Informatics, University of Edinburgh, Edinburgh, UK. [253]Department of Public Health, Ben Gurion University of the Negev, Beersheva, Israel. [254]Department of Physical Education, Federal University of Santa Catarina, Florianopolis, Brazil. [255]Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA. [256]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA. [257]Department of Epidemiology, School of Preventive Oncology, Patna, India. [258]Department of Epidemiology, Healis Sekhsaria Institute for Public Health, Mumbai, India. [259]Department of Nursing, Muhammadiyah University of Surakarta, Surakarta, Indonesia. [260]Department of Public Health, China Medical University, Taichung, Taiwan. [261]Department of Community Medicine, Ahmadu Bello University, Zaria, Nigeria. [262]Department of Medicine, University of Valencia, Valencia, Spain. [263]Carlos III Health Institute, Biomedical Research Networking Center for Mental Health Network (CIBERSAM), Madrid, Spain. [264]Department of Pediatrics, Hawassa University, Hawassa, Ethiopia. [265]International Vaccine Institute, Seoul, South Korea. [266]Department of Pediatrics, King Saud University, Riyadh, Saudi Arabia. [267]College of Medicine, Alfaisal University, Riyadh, Saudi Arabia. [268]Department of Anesthesiology, Perioperative, and Pain Medicine, Stanford University, Palo Alto, CA, USA. [269]Department of Anesthesiology, King Fahad Medical City, Riyadh, Saudi Arabia. [270]Department of Epidemiology and Biostatistics, University of Gondar, Gondar, Ethiopia. [271]Department of International Health, Johns Hopkins University, Baltimore, MD, USA. [272]Department of Pathology and Legal Medicine, University of São Paulo, Ribeirão Preto, Brazil. [273]Department of Health Economics, Hanoi Medical University, Hanoi, Vietnam. [274]Molecular Medicine and Pathology, University of Auckland, Auckland, New Zealand. [275]Clinical Hematology and Toxicology, Military Medical University, Hanoi, Vietnam. [276]Gomal Center of Biochemistry and Biotechnology, Gomal University, Dera Ismail Khan, Pakistan. [277]TB Culture Laboratory, Mufti Mehmood Memorial Teaching Hospital, Dera Ismail Khan, Pakistan. [278]Division of Health Sciences, University of Warwick, Coventry, UK. [279]Department of Epidemiology and Biostatistics, School of Public Health and Nutrition, Umeå University, Umeå, Sweden. [280]Department of Medical Mycology and Parasitology, Mazandaran University of Medical Sciences, Sari, Iran. [281]Argentine Society of Medicine, Ciudad De Buenos Aires, Argentina. [282]Velez Sarsfield Hospital, Buenos Aires, Argentina. [283]Psychosocial Injuries Research Center, Ilam University of Medical Sciences, Ilam, Iran. [284]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy. [285]Occupational Health Unit, Sant'orsola Malpighi Hospital, Bologna, Italy. [286]Department of Health Care Administration and Economics, National Research University Higher School of Economics, Moscow, Russia. [287]Foundation University Medical College, Foundation University Islamabad, Islamabad, Pakistan. [288]Department of Epidemiology and Biostatistics, Wuhan University, Wuhan, China. [289]Department of Psychiatry, University of São Paulo, São Paulo, Brazil. [290]University of Nairobi, Nairobi, Kenya. [291]School of Medicine, Nanjing University, Nanjing, China. [292]Department of Diabetes and Metabolic Diseases, University of Tokyo, Tokyo, Japan. [293]School of Allied Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia. [294]Department of Psychopharmacology, National Center of Neurology and Psychiatry, Tokyo, Japan. [295]Health Economics & Finance, Jackson State University, Jackson, MS, USA. [296]School of Medicine, Tsinghua University, Beijing, China. [297]Prevention of Cardiovascular Disease Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [298]Global Health Institute, Wuhan University, Wuhan, China. [299]Department of Medicine, School of Clinical Sciences at Monash Health, Monash University, Melbourne, Victoria, Australia. [300]Maternal and Child Health Division, International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh. [301]Student Research Committee, Babol University of Medical Sciences, Babol, Iran. [302]School of Public Health, Wuhan University of Science and Technology, Wuhan, China. [303]Hubei Province Key Laboratory of Occupational Hazard Identification and Control, Wuhan University of Science and Technology, Wuhan, China. [304]Department of Anesthesiology & Pain Medicine, University of Washington, Seattle, WA, USA. [305]These authors jointly supervised this work: Nicholas J. Kassebaum, Simon I. Hay. *e-mail: sihay@uw.edu

Methods

Overview

Building from our previous study of CGF in Africa[9], we used Bayesian model-based geostatistics[14]—which leveraged geo-referenced survey data and environmental and socioeconomic covariates, and the assumption that points with similar covariate patterns and that are closer to one another in space and time would be expected to have similar patterns of CGF—to produce high-spatial-resolution estimates of the prevalence of stunting, wasting, and underweight among children under five across LMICs. Stunting, wasting, and underweight were defined as *z*-scores that were two or more standard deviations below the WHO healthy population reference median for length/height-for-age, weight-for-length/height, and weight-for-age, respectively, for age- and sex-specific curves[6]. Using an ensemble modelling framework that feeds into a Bayesian generalized linear model with a correlated space–time error, and 1,000 draws from the fitted posterior distribution, we generated estimates of annual prevalence for each indicator of CGF on a 5 × 5-km grid over 105 LMICs for each year from 2000 to 2017 and mapped results at administrative levels to provide relevant subnational information for policy planning and public health action. For this analysis, we compiled an extensive geo-positioned dataset, using data from 460 household surveys and reports representing 4.6 million children. To ensure comparability with national estimates and to facilitate benchmarking, these local-level estimates were calibrated to those produced by the Global Burden of Disease (GBD) Study 2017[1], and were subsequently aggregated to the first administrative level (for example, states or provinces) and second administrative level (for example, districts or departments) in each LMIC. We also predict CGF prevalence for 2025 based on 2000–2017 trajectories and estimate the AROC required to meet the WHO GNTs by 2025. In addition, we estimate the 2017 absolute numbers of children under five affected by each CGF indicator in LMICs based on our prevalence estimates and the size of the populations of children under five[15,16]. Furthermore, we provide figures that demonstrate subnational disparities between each country's second administrative-level units with the highest and lowest estimated prevalence for 2000 and 2017 (Extended Data Figs. 2, 4, 6). We re-estimate CGF prevalence for the 51 African countries included in our previous analysis[9] using 28 additional surveys, and extend time trends to model each year from 2000 to 2017. Owing to these improvements in data availability and methodology, the estimates provided here supersede our previous modelling efforts.

Countries were selected for inclusion in this study using the socio-demographic index (SDI)—a summary measure of development that combines education, fertility, and poverty, published in the GBD study[1]. The analyses reported here include countries in the low, low-middle, and middle SDI quintiles, with several exceptions (Supplementary Table 3). China, Iran, Libya, and Malaysia were included despite high-middle SDI status in order to create better geographical continuity. Albania and Moldova were excluded owing to geographical discontinuity with other included countries and lack of available survey data. We did not estimate for the island nations of American Samoa, Federated States of Micronesia, Fiji, Kiribati, Marshall Islands, North Korea, Samoa, Solomon Islands, or Tonga, where no available survey data could be sourced. The flowchart of our modelling process is provided in Extended Data Fig. 9.

Surveys and child anthropometry data

We extracted individual-level height, weight, and age data for children under five from household survey series including the Demographic and Health Surveys (DHS), Multiple Indicator Cluster Surveys (MICS), Living Standards Measurement Study (LSMS), and Core Welfare Indicators Questionnaire (CWIQ), among other country-specific child health and nutrition surveys[7,17–19] (Supplementary Tables 4, 5). Included in our models were 460 geo-referenced household surveys and reports from 105 countries representing approximately 4.6 million children under five. Each individual child record was associated with a cluster, a group of neighbouring households or a 'village' that acts as a primary sampling unit. Some surveys included geographical coordinates or precise place names for each cluster within that survey (138,938 clusters for stunting, 144,460 for wasting, and 147,624 for underweight). In the absence of geographical coordinates for each cluster, we assigned data to the smallest available administrative areal unit in the survey (termed a 'polygon') while correcting for the survey sample design (16,554 polygons for stunting, 18,833 for wasting, and 19,564 for underweight). Boundary information for these administrative units was obtained as shapefiles either directly from the surveys or by matching to shapefiles in the Global Administrative Unit Layers (GAUL)[20] or the Database of Global Administrative Areas (GADM)[21]. In select cases, shapefiles provided by the survey administrator were used, or custom shapefiles were created based on survey documentation. These areal data were resampled to point locations using a population-weighted sampling approach over the relevant areal unit with the number of locations set proportionally to the number of grid cells in the area and the total weights of all the resampled points summing to one[16].

Select data sources were excluded for the following reasons: missing survey weights for areal data, missing sex variable, insufficient age granularity (in months) for calculations of length/height-for-age *z*-scores and weight-for-age *z*-scores in children ages 0–2 years, incomplete sampling (for example, only children ages 0–3 years measured), or untrustworthy data (as determined by the survey administrator or by inspection). We excluded data for children for whom we could not compute age in both months and weeks. Children with height values ≤0 cm or ≥180 cm, and/or with weight values ≤0 kg or ≥45 kg were also excluded from the study. We also excluded data that were considered outliers according to the 2006 WHO Child Growth Standards recommended range values, which were values <−6 or >6 length/height-for-age *z*-score for stunting, <−5 or >5 weight-for-length/height *z*-score for wasting, and <−6 or >5 weight-for-age *z*-score for underweight[3,4]. Details on the survey data excluded for each country are provided in Supplementary Table 6. Data availability plots for all the CGF indicators by country, type, and year are included in Supplementary Figs. 2–16.

Child anthropometry

Using the height, weight, age, and sex data for each individual, height-for-age, weight-for-height, and weight-for-age *z*-scores were calculated using the age-, sex-, and indicator-specific LMS (lambda-mu-sigma) values from the 2006 WHO Child Growth Standards[3,4]. The LMS methodology allows for Gaussian *z*-score calculations and comparisons to be applied to skewed, non-Gaussian distributions[22]. We classified stunting, wasting, or underweight if the height/length-for-age, weight-for-height/length, or weight-for-age, respectively, was more than two standard deviations (*z*-scores) below the WHO growth reference population[6]. These individual-level data observations were then collapsed to cluster-level totals for the number of children sampled and total number of children under five affected by stunting, wasting, or underweight.

Temporal resolution

We estimated the prevalence of stunting, wasting, and underweight annually from 2000 to 2017 using a model that allows us to account for data points measured across survey years. As such, the model would also allow us to predict at monthly or finer temporal resolutions; however, we are limited both computationally and by the temporal resolution of the covariates.

Seasonality adjustment

Owing to the acute nature of wasting and its relative temporal transience, wasting data were pre-processed to account for seasonality within each year of observation. Across LMICs, large proportions of the population live in rural areas and have livelihoods that rely on agriculture

and livestock. Seasonality affects the availability of and access to food, sometimes owing to natural disasters or climate events (for example, floods, monsoons, or droughts) that vary by season. Generalized additive models were fit to wasting data across time using the month of interview and a country-level fixed effect as the explanatory variables, and the wasting *z*-score as the response. A 12-month periodic spline for the interview month was used, as well as a spline that smoothed across the whole duration of the dataset. Once the models were fit, individual weight-for-height/length *z*-score observations were adjusted so that each measurement was consistent with a day that represented a mean day in the periodic spline. The seasonality adjustment had relatively little effect on the raw data[9].

Spatial covariates

To leverage strength from locations with observations to the entire spatiotemporal domain, we compiled several 5 × 5-km raster layers of possible socioeconomic and environmental correlates of CGF in the 105 LMICs (Supplementary Table 7, Supplementary Fig. 17). Covariates were selected based on their potential to be predictive for the set of CGF indicators, after reviewing literature on evidence and plausible hypotheses as to their influence. Acquisition of temporally dynamic datasets, where possible, was prioritized to best match our observations and thus predict the changing dynamics of the CGF indicators. Of the twelve covariates included, eight were temporally dynamic and were reformatted as a synoptic mean over each estimation period or as a mid-period year estimate: these covariates included average daily mean rainfall (precipitation), average daily mean temperature, enhanced vegetation index, fertility, malaria incidence, educational attainment in women of reproductive age (15–49 years old), population, and urbanicity. The remaining four covariate layers were static throughout the study period and were applied uniformly across all modelling years; growing season length, irrigation, nutritional yield for vitamin A, and travel time to nearest settlement of >50,000 inhabitants.

To select covariates and capture possible nonlinear effects and complex interactions between them, an ensemble covariate modelling method was implemented[23]. For each region, three sub-models were fit to our dataset using all of our covariate data as explanatory predictors; these sub-models were: generalized additive models, boosted regression trees, and lasso regression. Each sub-model was fit using fivefold cross-validation to avoid overfitting, and the out-of-sample predictions from across the five holdouts were compiled into a single comprehensive set of predictions from that model. In addition, the same sub-models were run using 100% of the data, and a full set of in-sample predictions were created. The three sets of out-of-sample sub-model predictions were fed into the full geostatistical model[14] as the explanatory covariates when performing the model fit. The in-sample predictions from the sub-models were used as the covariates when generating predictions using the fitted full geostatistical model. A recent study demonstrated that this ensemble approach can improve predictive validity by up to 25% over an individual model[23].

Geostatistical model analysis

Binomial count data were modelled within a Bayesian hierarchical modelling framework using a logit link function and a spatially and temporally explicit hierarchical generalized linear regression model to fit prevalence of each of our indicators in 14 regions[24] of LMICs (North Africa, western SSA, central SSA, eastern SSA, southern SSA, Middle East, Central Asia, East Asia, South Asia, Southeast Asia, Oceania, Central America and the Caribbean, Andean South America, and Tropical South America; see Extended Data Fig. 10). For each region, we explicitly wrote the hierarchy that defines our Bayesian model.

For each binomial CGF indicator, we modelled the average number of children with stunting, wasting, or who were underweight in each survey cluster, d. Survey clusters are precisely located by their GPS coordinates and year of observation, which we map to a spatial raster location, i, at time, t. We observed the number of children reported to be stunted, wasted, or underweight, respectively, as binomial count data, C_d, among an observed sample size, N_d. As we may have observed several data clusters within a given location, i, at time, t, we refer to the probability of stunting, wasting, or underweight, p, within a given cluster, d, by its indexed location, i, and time, t, as $p_{i(d),t(d)}$.

$$C_d|p_{i(d),t(d)},\ N_d \sim \text{Binomial}(p_{i(d),t(d)},\ N_d)\ \forall\ \text{observed clusters}\ d$$

$$\text{logit}(p_{i,t}) = \beta_0 + \mathbf{X}_{i,t}\boldsymbol{\beta} + Z_{i,t} + \epsilon_{\text{ctr}(i)} + \epsilon_{i,t} + Z_{i,t}\ \forall$$
$$i \in \text{spatial domain}\ \forall\ t \in \text{time domain}$$

$$\sum_{h=1}^{3} \beta_h = 1$$
$$\epsilon_{\text{ctr}} \sim \text{iid Normal}(0,\ \gamma^2)$$
$$\epsilon_{i,t} \sim \text{iid Normal}(0,\ \sigma^2)$$
$$\mathbf{Z} \sim \text{GP}(0, \Sigma^{\text{space}} \otimes \Sigma^{\text{time}})$$
$$\Sigma^{\text{space}} = \frac{\omega^2}{\Gamma(\nu)2^{\nu-1}} \times (\kappa D)^{\nu} \times K_{\nu}(\kappa D)$$
$$\Sigma^{\text{time}}_{j,k} = \rho^{|k-j|}$$

For indices d, i, and t, *(index) is the value of * at that index. The probabilities, $p_{i,t}$, represent both the annual prevalence at the space–time location and the probability that an individual child was afflicted with the risk factor given that they lived at that particular location. The annual prevalence, $p_{i,t}$, of each indicator was modelled as a linear combination of the three sub-models (generalized additive model, boosted regression trees, and lasso regression), rasterized covariate values, $\mathbf{X}_{i,t}$, a correlated spatiotemporal error term, $Z_{i,t}$, and country random effects, $\epsilon_{\text{ctr}(i)}$, with one unstructured country random effect fit for each country in the modelling region and all ϵ_{ctr} sharing a common variance parameter, γ^2, and an independent nugget effect, $\epsilon_{i,t}$, with variance parameter, σ^2. Coefficients in β_h in the three sub-models $h = 1, 2, 3$ represent their respective predictive weighting in the mean logit link, while the joint error term, $Z_{i,t}$, accounts for residual spatiotemporal autocorrelation between individual data points that remains after accounting for the predictive effect of the sub-model covariates, the country-level random effect, $\epsilon_{\text{ctr}(i)}$, and the nugget independent error term, $\epsilon_{i,t}$. The residuals, $Z_{i,t}$, are modelled as a three-dimensional Gaussian process (GP) in space–time centred at zero and with a covariance matrix constructed from a Kronecker product of spatial and temporal covariance kernels. The spatial covariance, Σ^{space}, is modelled using an isotropic and stationary Matérn function[25], and temporal covariance, Σ^{time}, as an annual autoregressive (AR1) function over the 18 years represented in the model. In the stationary Matérn function, Γ is the gamma function, K_{ν} is the modified Bessel function of order $\nu > 0$, $\kappa > 0$ is a scaling parameter, D denotes the Euclidean distance, and ω^2 is the marginal variance. The scaling parameter, κ, is defined to be $\kappa = \sqrt{8\nu}/\delta$ in which δ is a range parameter (which is about the distance where the covariance function approaches 0.1) and ν is a scaling constant, which is set to 2 rather than fit from the data[26,27]. This parameter is difficult to reliably fit, as documented by many other analyses[26,28,29] that set this to 2. The number of rows and the number of columns of the spatial Matérn covariance matrix are both equal to the number of spatial mesh points for a given modelling region. In the AR1 function, ρ is the autocorrelation function (ACF), and k and j are points in the time series where $|k-j|$ defines the lag. The number of rows and the number of columns of the AR1 covariance matrix are both equal to the number of temporal mesh points (18). The number of rows and the number of columns of the space–time covariance matrix, $\Sigma^{\text{space}} \otimes \Sigma^{\text{time}}$, for a given modelling region are both equal to: (the number of spatial mesh points × the number of temporal mesh points).

This approach leveraged the residual correlation structure of the data to more accurately predict prevalence estimates for locations with no data, while also propagating the dependence in the data through to uncertainty estimates[14]. The posterior distributions were fit using computationally efficient and accurate approximations in R-INLA[30,31] (integrated nested Laplace approximation) with the stochastic partial differential equations (SPDE)[27] approximation to the Gaussian process residuals using R project v.3.5.1. The SPDE approach using INLA has been demonstrated elsewhere, including the estimation of health indicators, particulate air matter, and population age structure[9,32–35]. Uncertainty intervals were generated from 1,000 draws (that is, statistically plausible candidate maps)[36] created from the posterior-estimated distributions of modelled parameters. Further details on model and estimation processes are provided in the Supplementary Information.

Post estimation

To leverage national-level data included in the 2017 GBD study[1] that were not within the scope of our current geospatial modelling framework, and to ensure alignment between these estimates and GBD national-level and subnational estimates, we performed a post hoc calibration to the mean of the 1,000 draws. We calculated population-weighted aggregations to the GBD estimate level, which was either at the national or first administrative level, and compared these estimates to our corresponding year estimates from 2000 to 2017. We defined the calibration factor to be the ratio between the GBD estimates and our current estimates for each year from 2000 to 2017. For some selected countries where GBD estimates were at the first administrative level, the calibration factors were also calculated at the lowest available subnational level. These countries included Brazil, China, Ethiopia, India, Indonesia, Iran, Mexico, and South Africa. Finally, we multiplied each of our estimates in a country-year (or first-administrative-year) by its associated factor. This ensures consistency between our geospatial estimates and those of the 2017 GBD[1], while preserving our estimated within-country geospatial and temporal variation. To transform grid-cell-level estimates into a range of information useful to a wide constituency of potential users, these estimates were aggregated at first and second administrative-level units specific to each country and at national levels using conditional simulation[37].

Although the models can predict all locations covered by available raster covariates, all final model outputs for which land cover was classified as 'barren or sparsely vegetated' on the basis of the most recently available Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data (2013) were masked[38]. Areas where the total population density was less than ten individuals per 1 × 1-km grid cell were also masked in the final outputs.

Model validation

We assessed the predictive performance of the models using fivefold out-of-sample cross-validation strategies and found that our prevalence estimates closely matched the survey data. To offer a more stringent analysis by respecting some of the spatial correlation in the data, holdout sets were created by combining sets of data at different spatial resolutions (for example, first administrative level). Validation was performed by calculating bias (mean error), variance (root mean square error), 95% data coverage within prediction intervals, and correlation between observed data and predictions. All validation metrics were calculated on the out-of-sample predictions from the fivefold cross-validation. Furthermore, measures of spatial and temporal autocorrelation pre- and post-modelling were examined to verify correct recognition, fitting, and accounting for the complex spatiotemporal correlation structure in the data. All validation procedures and corresponding results are included in Supplementary Tables 14–22 and Supplementary Figs. 24–41.

Projections

To compare our estimated rates of improvement in CGF prevalence over the last 18 years with the improvements needed between 2017 and 2025 to meet WHO GNTs, we performed a simple projection using estimated annualized rates of change (AROC) applied to the final year of our estimates.

For each CGF indicator, u, we calculated AROC at each grid cell, m, by calculating the AROC between each pair of adjacent years, t:

$$\mathrm{AROC}_{u,m,t} = \mathrm{logit}\left(\frac{p_{u,m,t}}{p_{u,m,t-1}}\right)$$

We then calculated a weighted AROC for each indicator by taking a weighted average across the years, where more recent AROCs were given more weight in the average. We defined the weights to be:

$$W_t = (t - 2000 + 1)^{\gamma}$$

in which γ may be chosen to give varying amounts of weight across the years. For any indicator, we then calculated the average AROC to be:

$$\mathrm{AROC}_{u,m} = \mathrm{logit}\left(\sum_{2001}^{2017} W_t \times \mathrm{AROC}_{u,m,t}\right)$$

Finally, we calculated the projections, Proj, by applying the AROC in our 2017 mean prevalence estimates to produce estimates in 8 years from 2017 to 2025. For this set of projections, we selected $\gamma = 1.7$ for stunting, $\gamma = 1.9$ for wasting, and $\gamma = 1.8$ for underweight[1].

$$\mathrm{Proj}_{u,m,2025} = \mathrm{logit}^{-1}(\mathrm{logit}(p_{u,m,2017}) + \mathrm{AROC}_{u,m} \times 8)$$

This projection scheme is analogous to the methods used in the 2017 GBD measurement of progress and projected attainment of health-related Sustainable Development Goals[1]. Our projections are based on the assumption that areas will sustain the current AROC, and the precision is dependent on the level of uncertainty emanating from the estimation of annual prevalence.

Although the WHO GNT for wasting was to reduce prevalence to less than 5%, the WHO GNT for stunting was a 40% relative reduction in prevalence. For our analyses, we defined the WHO GNT for stunting and underweight (for which no WHO GNT was established) to be 40% reduction relative to 2010, the year the World Health Assembly requested the development of the WHO GNTs[39].

Limitations

The accuracy of our models depends on the volume, representativeness, quality, and validity of surveys available for analysis (Supplementary Tables 4, 5, Supplementary Figs. 2–16). Persistent data gaps in national surveys include a lack of CGF data or household-level characteristics, such as hygiene and sanitation practices. The associated uncertainties of our estimates are higher in areas where data are either missing or less reliable (Figs. 1d, 2d, Extended Data Fig. 5d), and rely more heavily on covariates and borrowing from neighbouring areas for their modelling (Supplementary Table 7, Supplementary Fig. 17). Investments in improvements of health surveillance systems and including child anthropometrics as part of routine data collection for profiling population characteristics could improve the certainty of our estimates and better monitor progress towards international goals. In addition, measurement error in collecting anthropometric information, including the child's age, height, and weight, could have introduced bias or error in the data across different survey types. The accuracy of age data may be affected by differences in sampling approaches and self-reporting bias, such as long recall period or selective recall. Weight and height measurements may be inaccurate owing to improper calibration

of equipment, device inaccuracy, different measurement methods, or human error. We did not include a survey random effect to account for between-survey variability in data accuracy; given that most surveys represent a country-year, it would be difficult to distinguish these biases from temporal effects. Our calibration approach in the post-estimation process used only a ratio estimator and did not account for an additive effect, which may have introduced bias. Owing to the complexity of the boosted regression tree sub-model, we were unable to account for the uncertainty of our three sub-models in our final estimates (see Supplementary Information section 3.2.2 for more detail). It is worth noting that our analyses are descriptive and do not support causal inferences on their own. Future research is required to determine the causal pathways for each CGF indicator across and within LMICs.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

CGF estimates can be further explored at various spatial scales (national, administrative, and local levels) through our customized online data visualization tools (https://vizhub.healthdata.org/lbd/cgf). The full output of the analyses and the underlying data used in the analyses are publicly available via the Global Health Data Exchange (GHDx; http://ghdx.healthdata.org/record/ihme-data/lmic-child-growth-failure-geospatial-estimates-2000-2017). Some data sources are under special licenses for the current study and are thus not publicly available. Supplementary Tables 4 and 5 show the incorporated data sources, and data with restrictions are marked with an obelisk symbol (†). All maps presented in this study are generated by the authors and no permissions are required to publish them.

The findings of this study are supported by data available in public online repositories, data publicly available upon request of the data provider, and data not publicly available owing to restrictions by the data provider. Non-publicly available data were used under license for the current study but may be available from the authors upon reasonable request and with permission of the data provider. Detailed tables and figures of data sources and availability can be found in Supplementary Tables 4, 5, and Supplementary Figs. 2–16.

Administrative boundaries were retrieved from the Global Administrative Unit Layers (GAUL)[20] or the Database of Global Administrative Areas (GADM)[21]. Land cover was retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota[40]. Lakes were retrieved from the Global Lakes and Wetlands Database (GLWD), courtesy of the World Wildlife Fund and the Center for Environmental Systems Research, University of Kassel[41,42]. Populations were retrieved from WorldPop[15,16]. All maps in this study were produced using ArcGIS Desktop 10.6.

Code availability

Our study follows the Guidelines for Accurate and Transparent Health Estimate Reporting (GATHER; Supplementary Table 1). All code used for these analyses is publicly available online http://ghdx.healthdata.org/record/ihme-data/lmic-child-growth-failure-geospatial-estimates-2000-2017 and at http://github.com/ihmeuw/lbd/tree/cgf-lmic-2019.

14. Hotez, P. J. & Ribeiro, P. J. *Model-Based Geostatistics* (Springer New York, 2007).
15. WorldPop. *WorldPop Dataset* http://www.worldpop.org.uk/data/get_data/ (accessed 24 July 2017).
16. Tatem, A. J. WorldPop, open data for spatial demography. *Sci. Data* **4**, 170004 (2017).
17. UNICEF. *Multiple Indicator Cluster Surveys (MICS)* http://mics.unicef.org (accessed 26 June 2019).
18. World Bank Group. *Living Standards Measurement Survey (LSMS)*. http://surveys.worldbank.org/lsms (accessed 26 June 2019).
19. World Bank Group. *Core Welfare Indicators Questionnaire Survey (CWIQ)* http://ghdx.healthdata.org/series/core-welfare-indicators-questionnaire-survey-cwiq (accessed 21 April 2017).
20. GeoNetwork. *The Global Administrative Unit Layers (GAUL)* http://www.fao.org/geonetwork/srv/en/main.home (2015).
21. Global Administrative Areas (GADM). *GADM Database of Global Administrative Areas* http://www.gadm.org (2018).
22. Indrayan, A. Demystifying LMS and BCPE methods of centile estimation for growth and other health parameters. *Indian Pediatr.* **51**, 37–43 (2014).
23. Bhatt, S. et al. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *J. R. Soc. Interface* **14**, 20170520 (2017).
24. Murray, C. J. L. et al. GBD 2010: design, definitions, and metrics. *Lancet* **380**, 2063–2066 (2012).
25. Stein, M. L. *Interpolation of Spatial Data* (Springer New York, 1999).
26. Lindgren, F. & Rue, H. Bayesian spatial modelling with R-INLA. *J. Stat. Softw.* **63**, jss.v063.i19 (2015).
27. Lindgren, F., Rue, H. & Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. Series B Stat. Methodol.* **73**, 423–498 (2011).
28. Rozanov, Y. A. *Markov Random Fields* (Springer-Verlag, 1982).
29. Whittle, P. On stationary processes in the plane. *Biometrika* **41**, 434–449 (1954).
30. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. Series B Stat. Methodol.* **71**, 319–392 (2009).
31. Martins, T. G., Simpson, D., Lindgren, F. & Rue, H. Bayesian computing with INLA: new features. *Comput. Stat. Data Anal.* **67**, 68–83 (2013).
32. Golding, N. et al. Mapping under-5 and neonatal mortality in Africa, 2000-15: a baseline analysis for the Sustainable Development Goals. *Lancet* **390**, 2171–2182 (2017).
33. Cameletti, M., Lindgren, F., Simpson, D. & Rue, H. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *AStA Adv. Stat. Anal.* **97**, 109–131 (2013).
34. Alegana, V. A. et al. Fine resolution mapping of population age-structures for health and development applications. *J. R. Soc. Interface* **12**, 20150073 (2015).
35. Kinyoki, D. K. et al. Assessing comorbidity and correlates of wasting and stunting among children in Somalia using cross-sectional household surveys: 2007 to 2010. *BMJ Open* **6**, e009854 (2016).
36. Patil, A. P., Gething, P. W., Piel, F. B. & Hay, S. I. Bayesian geostatistics in health cartography: the perspective of malaria. *Trends Parasitol.* **27**, 246–253 (2011).
37. Gething, P. W., Patil, A. P. & Hay, S. I. Quantifying aggregated uncertainty in *Plasmodium falciparum* malaria prevalence and populations at risk via efficient space-time geostatistical joint simulation. *PLOS Comput. Biol.* **6**, e1000724 (2010).
38. Scharlemann, J. P. W. et al. Global data for ecology and epidemiology: a novel algorithm for temporal Fourier processing MODIS data. *PLoS ONE* **3**, e1408 (2008).
39. de Onis, M. et al. The World Health Organization's global target for reducing childhood stunting by 2025: rationale and proposed actions. *Matern. Child Nutr.* **9**, 6–26 (2013).
40. Friedl, M. & Sulla-Menashe, D. MCD12Q1 v006. MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500m SIN Grid https://doi.org/10.5067/MODIS/MCD12Q1.006 (NASA EOSDIS Land Processes DAAC, 2019).
41. Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol. (Amst.)* **296**, 1–22 (2004).
42. World Wildlife Fund. *Global Lakes and Wetlands Database, Level 3* https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database (2004).

Acknowledgements This work was primarily supported by grant OPP1132415 from the Bill & Melinda Gates Foundation.

Author contributions S.I.H. and N.J.K. conceived and planned the study. B.V.P., A.L.-A., and D.K.K. obtained, extracted, processed, and geo-positioned CGF data. L.E. constructed covariate data layers. D.K.K., A.E.O.-Z., and M.L.C. wrote the computer code and designed the statistical analyses. D.K.K. carried out the statistical analyses with input from A.E.O.-Z., M.L.C., N.J.H., and N.V.B. D.K.K. and L.E. prepared figures. D.K.K., L.E.S., and L.B.M. wrote the first draft of the manuscript with assistance from S.I.H. and M.F.S., and all authors contributed to subsequent revisions. All authors provided intellectual input into aspects of this study. Additional details on author contributions can be found in the Supplementary Information (section 8.0).

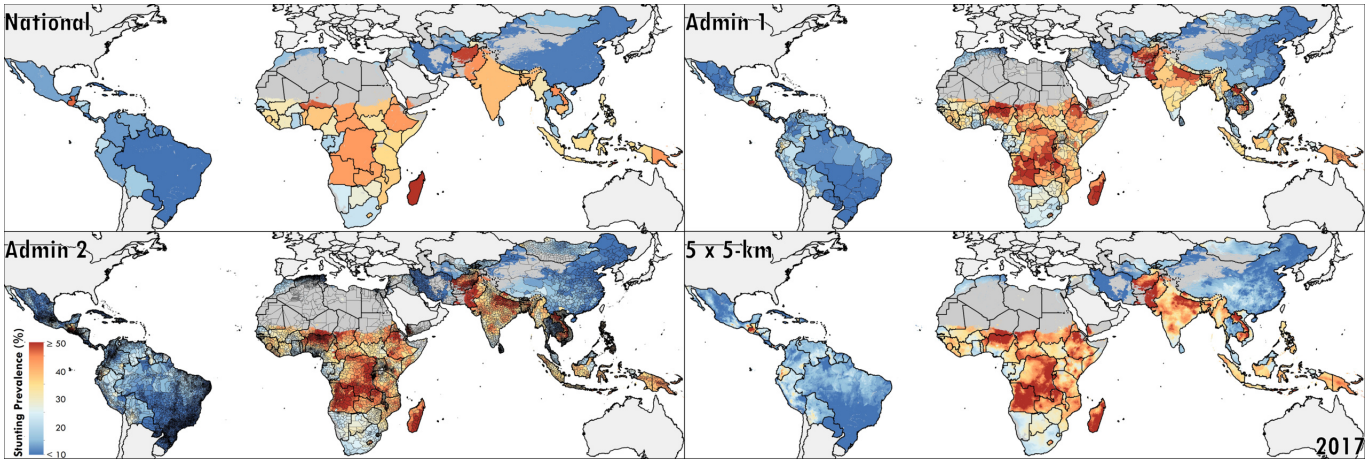
Competing interests This study was funded by the Bill & Melinda Gates Foundation. Co-authors employed by the Bill & Melinda Gates Foundation provided feedback on initial maps and drafts of this manuscript. Otherwise, the funders of the study had no role in study design, data collection, data analysis, data interpretation, writing of the final report, or the decision to publish. The corresponding author had full access to all the data in the study and had final responsibility for the decision to submit for publication.

Additional information

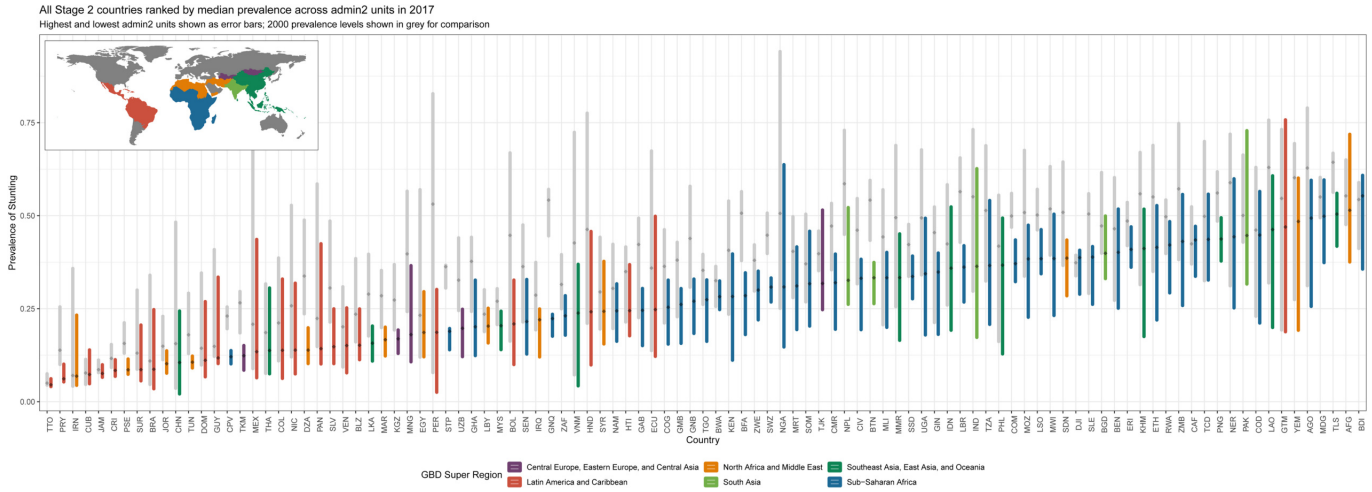
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1878-8.

Correspondence and requests for materials should be addressed to S.I.H.

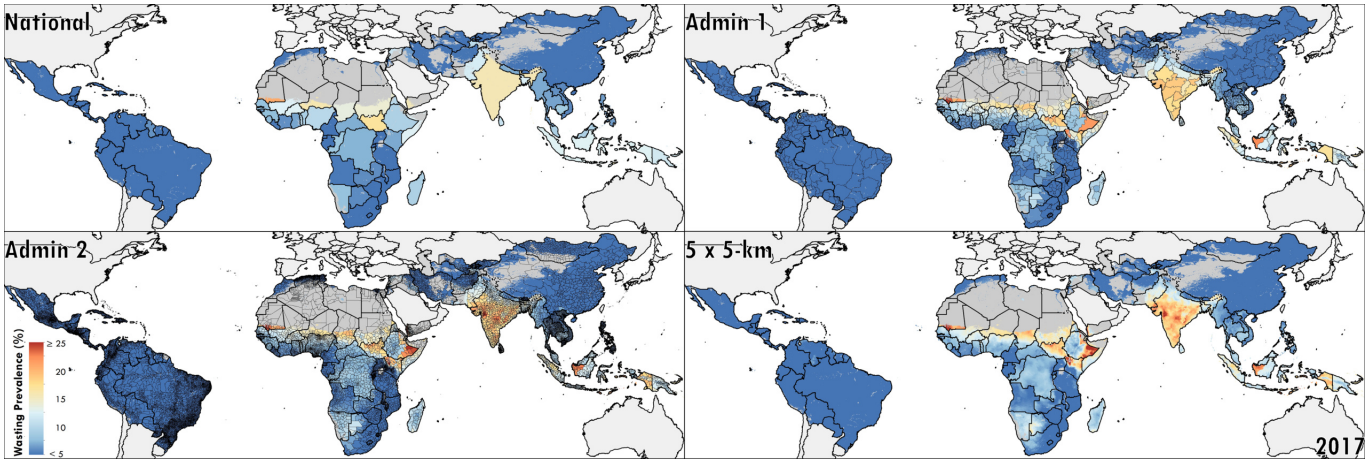
Reprints and permissions information is available at http://www.nature.com/reprints.



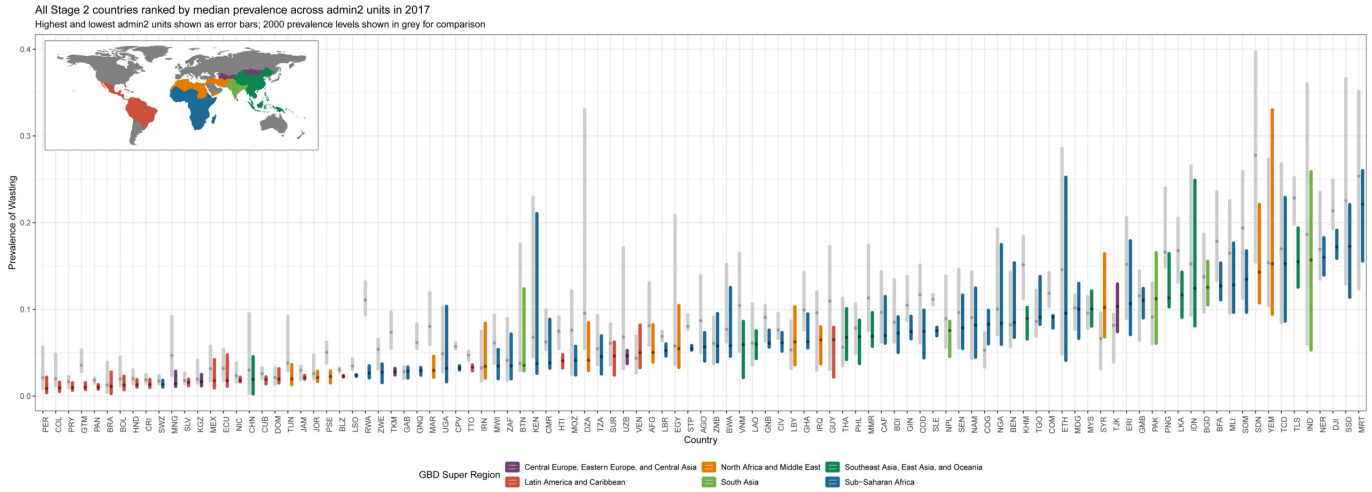
Extended Data Fig. 1 | Prevalence of stunting in children under five in LMICs at administrative levels 0, 1, 2, and at 5 × 5-km resolution in 2017. Administrative level 0 are national-level estimates; administrative level 1 are first administrative-level (for example, states or provinces) estimates; administrative level 2 are second administrative-level (for example, districts or departments) estimates. Maps reflect administrative boundaries, land cover, lakes, and population; grey-coloured grid cells had fewer than ten people per 1 × 1-km grid cell and were classified as 'barren or sparsely vegetated'[15,16,20,21,40–42], or were not included in these analyses. Maps were produced using ArcGIS Desktop 10.6.



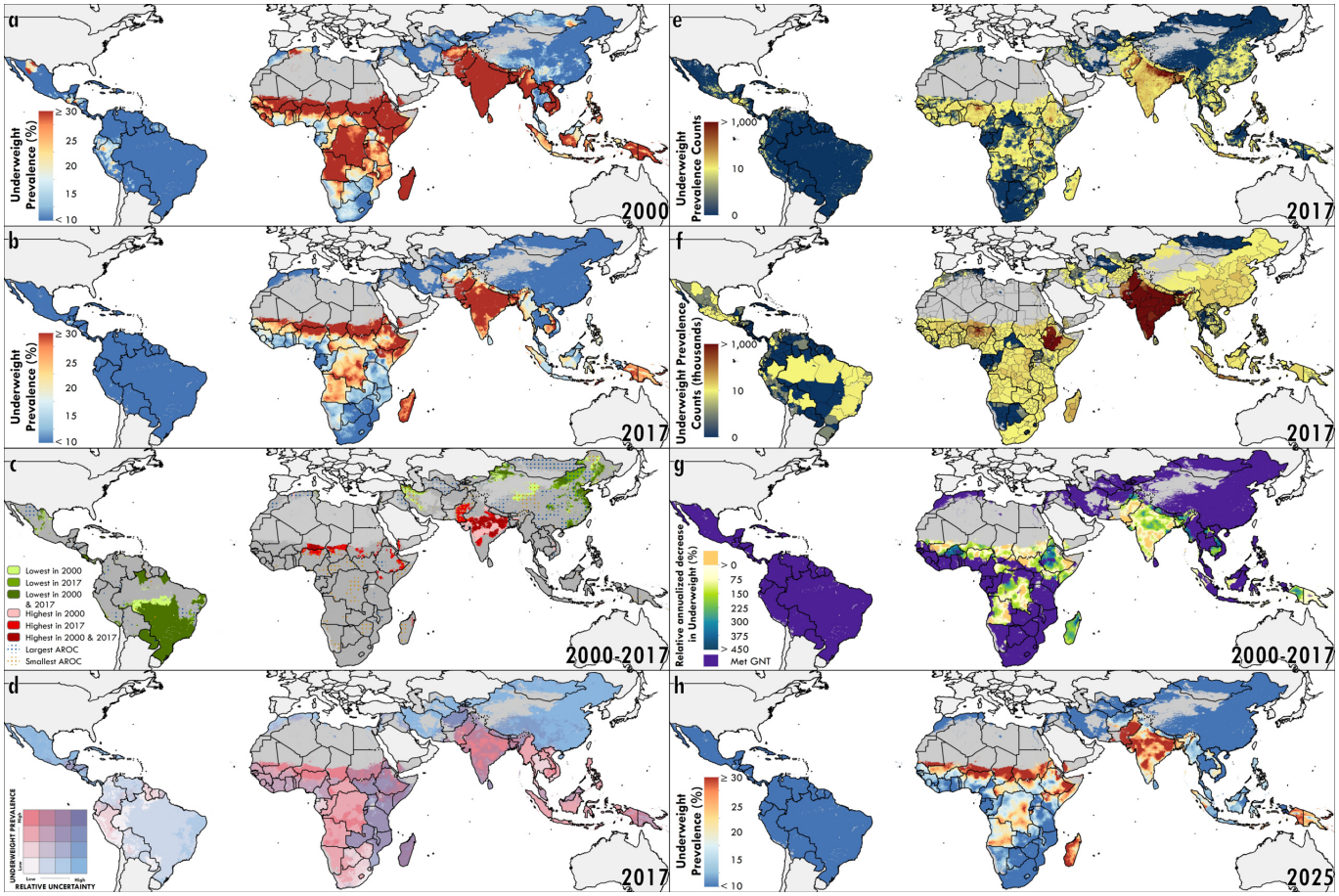
Extended Data Fig. 2 | Geographical inequality in the prevalence of child stunting across 105 countries. The bars represent the range of stunting prevalence in children under five in the second administrative-level units in each country. Bars indicating the range in 2017 are coloured according to the regions defined by the Global Burden of Disease (GBD)[1]. Grey bars indicate the range in 2000. The graph was produced using R project v.3.5.1.



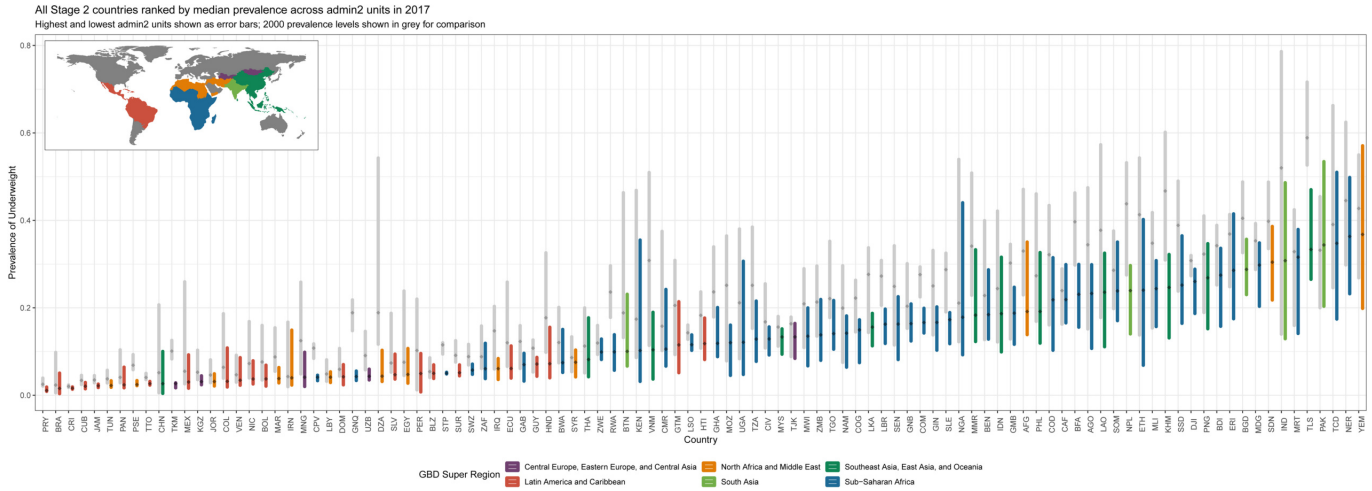
Extended Data Fig. 3 | Prevalence of wasting in children under five in LMICs at administrative levels 0, 1, 2, and at 5 × 5-km resolution in 2017. Administrative levels are as described in Extended Data Fig. 1. Maps reflect administrative boundaries, land cover, lakes, and population; grey-coloured grid cells had fewer than ten people per 1 × 1-km grid cell and were classified as 'barren or sparsely vegetated'[15,16,20,21,40–42], or were not included in these analyses. Maps were produced using ArcGIS Desktop 10.6.



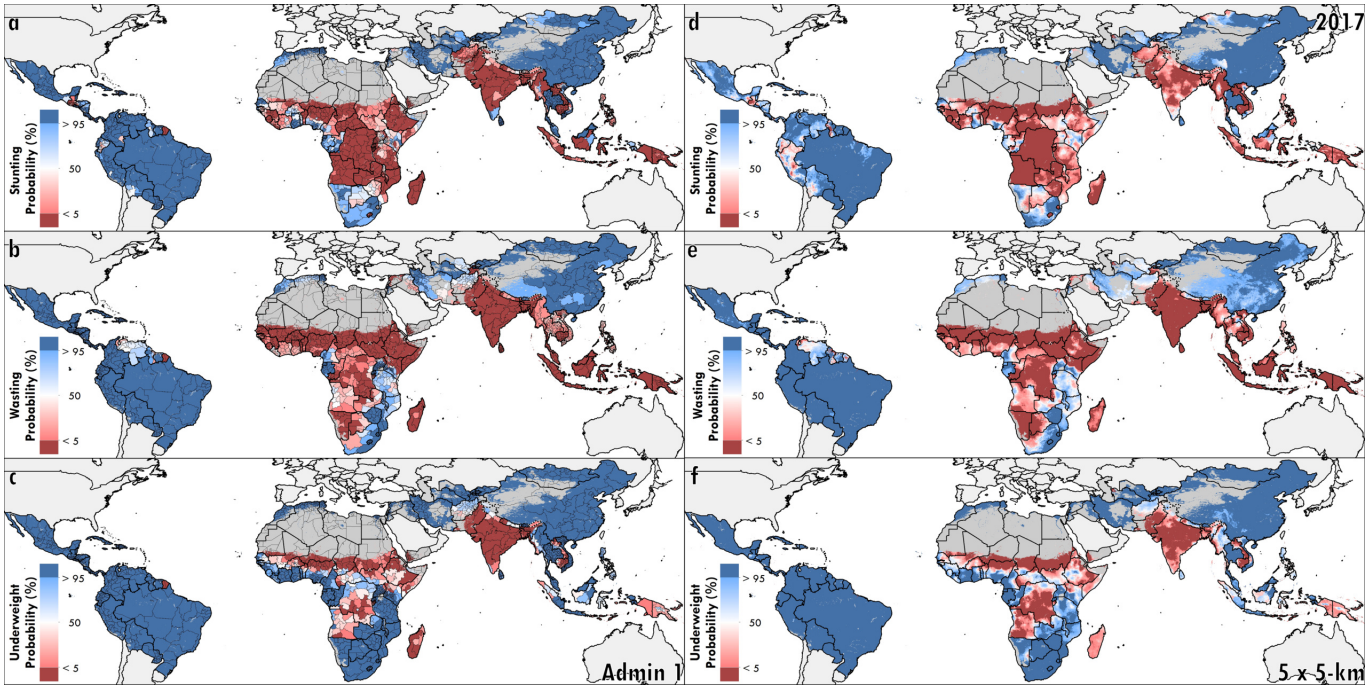
Extended Data Fig. 4 | Geographical inequality in prevalence of child wasting across 105 countries. The bars represent the range of wasting prevalence in children under five in the second administrative-level units in each country. Bars indicating the range in 2017 are coloured according to their GBD-defined[1] regions. Grey bars indicate the range in 2000. The graph was produced using R project v.3.5.1.



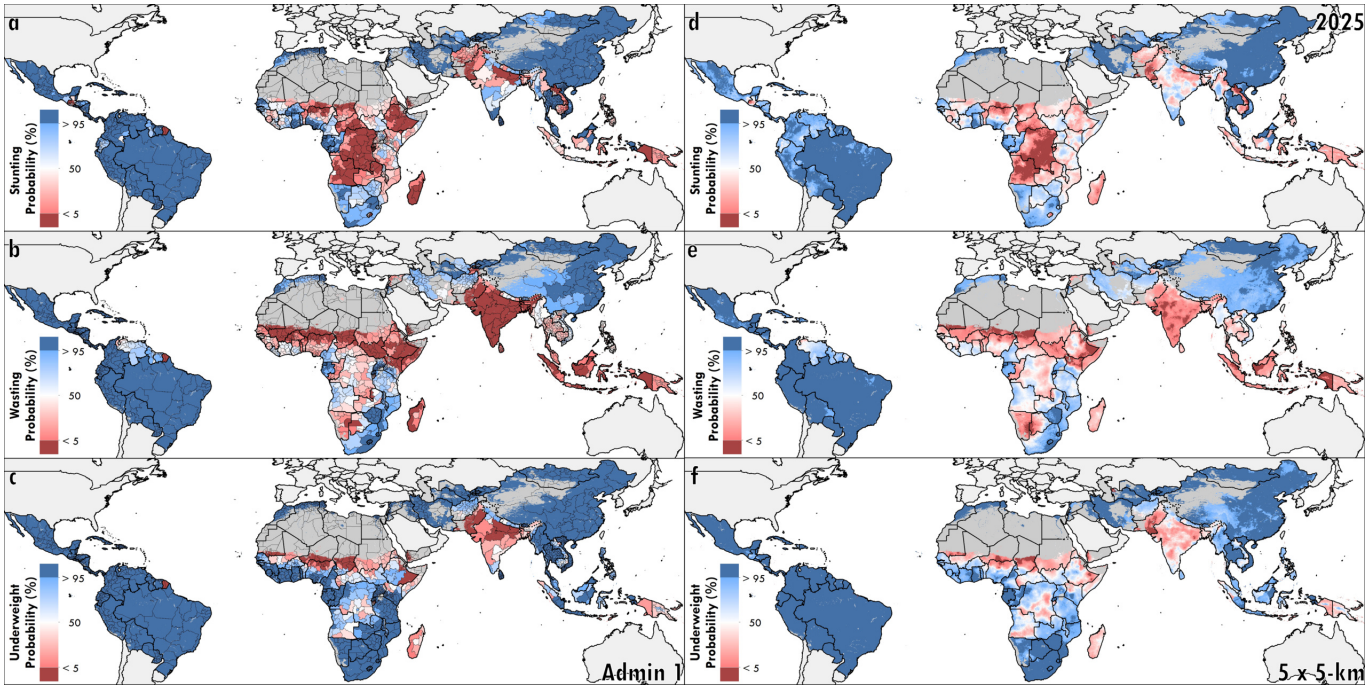
Extended Data Fig. 5 | Prevalence of underweight in children under five in LMICs (2000–2017) and progress towards 2025. a, **b**, Prevalence of underweight in children under five at the 5 × 5-km resolution in 2000 (**a**) and 2017 (**b**). **c**, Overlapping population-weighted tenth and ninetieth percentiles (lowest and highest) of 5 × 5-km grid cells and AROC in underweight, 2000–2017. **d**, Overlapping population-weighted quartiles of underweight prevalence and relative 95% uncertainty in 2017. **e**, **f**, Number of underweight children under five, at the 5 × 5-km (**e**) and first-administrative-unit (**f**) levels. **g**, 2000–2017 annualized decrease in underweight prevalence relative to rates needed during 2017–2025 to meet WHO GNT. **h**, Grid-cell-level predicted underweight prevalence in 2025. Maps were produced using ArcGIS Desktop 10.6. Interactive visualization tools are available at https://vizhub.healthdata.org/lbd/cgf.



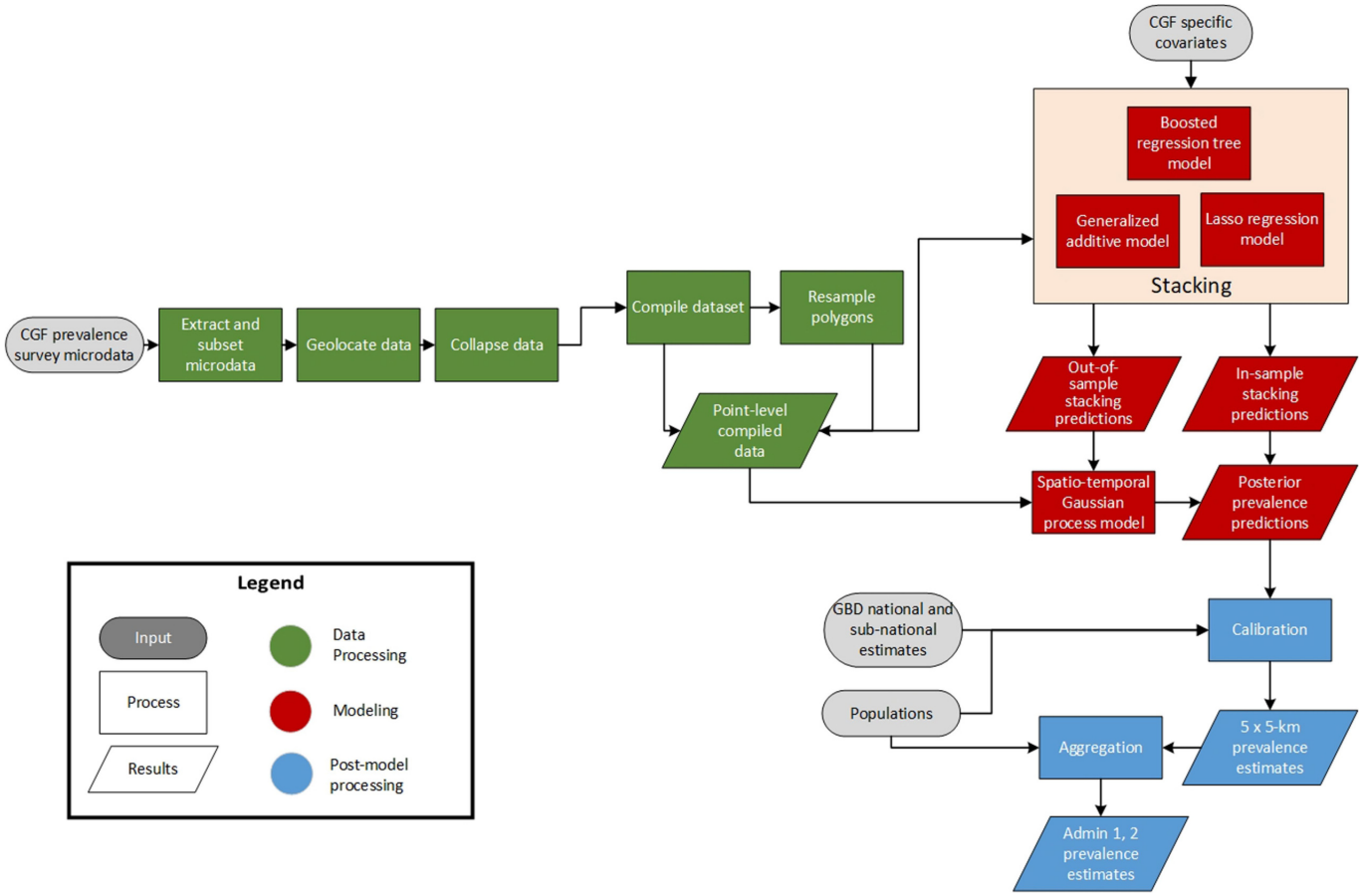
Extended Data Fig. 6 | Geographical inequality in prevalence of child underweight across 105 countries. The bars represent the range of underweight prevalence in the second administrative-level units in each country. Bars indicating the range in 2017 are coloured according to their GBD-defined[1] regions. Grey bars indicate the range in 2000. The graph was produced using R project v.3.5.1.



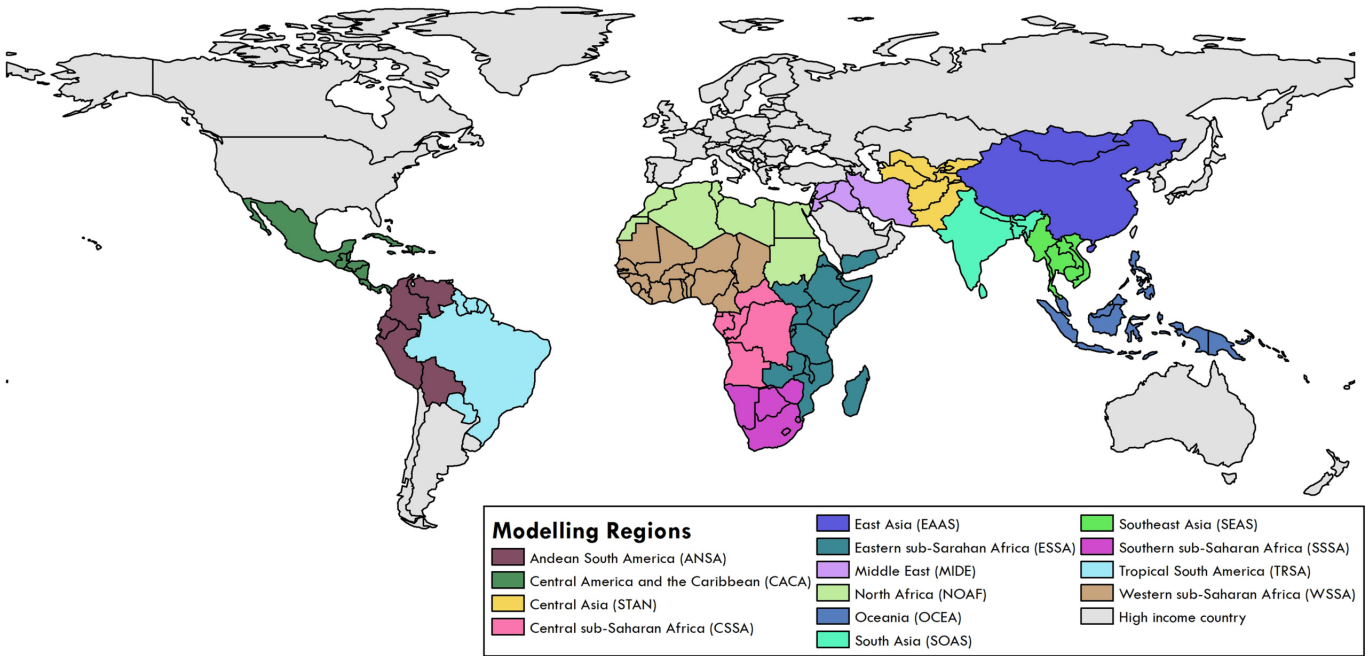
Extended Data Fig. 7 | Probability that WHO GNT had been achieved in 2017 at the first administrative and 5 × 5-km grid-cell levels for stunting, wasting, and underweight. a–**f**, Probability of WHO GNT achievement in 2017 at the first administrative and 5 × 5-km levels for stunting (**a**, **d**), wasting (**b**, **e**), and underweight (**c**, **f**). Dark-blue and dark-red grid cells indicate >95% and <5% probability, respectively, of having met the WHO GNT in 2017. Given that there was no WHO GNT established for underweight, we based the underweight target on WHO GNT for stunting, as the conditions are similarly widespread and prevalent. Maps were produced using ArcGIS Desktop 10.6.



Extended Data Fig. 8 | Probability of meeting WHO GNT in 2025 at the first administrative and 5 × 5-km grid-cell levels for stunting, wasting, and underweight. a–**f**, Probability of WHO GNT achievement in 2025 at the first administrative and 5 × 5-km levels for stunting (**a**, **d**), wasting (**b**, **e**), and underweight (**c**, **f**). Dark-blue and dark-red grid cells indicate >95% and <5% probability, respectively, of meeting WHO GNT in 2025. Given that there was no WHO GNT established for underweight, we based the underweight target on WHO GNT for stunting as the conditions are similarly widespread and prevalent. Maps were produced using ArcGIS Desktop 10.6.



Extended Data Fig. 9 | Flowchart of CGF prevalence modelling process. The process used to produce CGF prevalence estimates in LMICs involved three main parts. In the data-processing steps (green), data were identified, extracted, and prepared for use in the models. In the modelling phase (red), we used these data and covariates in stacked generalization ensemble models and spatiotemporal Gaussian process models for each CGF indicator. In post-processing (blue), we calibrated the prevalence estimates to match 2017 GBD study[1] estimates and aggregated the estimates to the first- and second-administrative-level units in each country.



Extended Data Fig. 10 | Modelling regions. Modelling regions[24] were based on geographical and SDI regions from the GBD study[1], defined as: Andean South America, Central America and the Caribbean, central SSA, East Asia, eastern SSA, Middle East, North Africa, Oceania, Southeast Asia, South Asia, southern SSA, Central Asia, Tropical South America, and western SSA. 'High income country' refers to regions not included in our models owing to high-middle or a high SDI. The map was produced using ArcGIS Desktop 10.6.

nature research

Corresponding author(s): Simon I. Hay

Last updated by author(s): 2019/10/18

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No primary data collection was carried out for these analyses.
Data analysis	These analyses were carried out using R version 3.5.1. The main geostatistical models were fit using R-INLA version 18.07.12. All code used for these analyses is publicly available online at http://github.com/ihmeuw/lbd/tree/cgf-lmic-2019

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The findings of this study are supported by data that are available in public online repositories, data that are publicly available upon request from the data provider, and data that are not publicly available due to restrictions by the data provider and which were used under license for the current study. Detailed tables of data sources can be found in Supplementary Tables 2-6. More information about each data source is available on the Global Health Data Exchange (http://ghdx.healthdata.org/), including information about the data provider and links to where the data can be accessed or requested (where available).

Administrative boundaries were retrieved from the Global Administrative Unit Layers (GAUL) dataset, implemented by FAO within the CountrySTAT and Agricultural Market Information System (AMIS) projects[16]. Land cover was retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota[18]. Lakes were retrieved from the Global Lakes and Wetlands Database (GLWD), courtesy of the World Wildlife Fund and the Center for Environmental Systems Research, University of Kassel[20].

Populations were retrieved from WorldPop[14,15].

All estimates produced as part of these analyses are publicly available from the Global Health Data Exchange (http://ghdx.healthdata.org/record/ihme-data/lmic-child-growth-failure-geospatial-estimates-2000-2017) and via a user-friendly data visualisation tool (https://vizhub.healthdata.org/lbd/cgf)

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The sample size is reported in the methods section. The final database for stunting consists of 142,468 clusters and 18,010 polygons with a sample size totaling over 3.9 million children in LMICs. The database for wasting consists of 142,017 clusters and 17,997 polygons with a sample size totaling over 3.9 million children in LMICs. The database for underweight consists of 142,528 clusters and 18,192 polygons with a sample size of over 4.0 million children in LMICs. In total, 460 household surveys and reports were collected representing 4.6 million children. This is an observational study with no hypothesis testing and the sample size was not pre-specified. We evaluate the overall performance of our modelling strategy, given the available data, as part of a validation exercise as described in the 'Model validation' section of the methods, and as reported in the Supplementary Information (Supplementary Table 9).
Data exclusions	Select data sources were excluded for the following reasons: missing survey weights for areal data, missing sex variable, insufficient age granularity (in months) for HAZ and WAZ calculations in children ages 0–2 years, incomplete sampling (e.g., only children ages 0–3 years measured), or untrustworthy data (as determined by the survey administrator or by inspection). We excluded data for children for whom we could not compute age in both months and weeks. Children with height values ≤0 cm or ≥180 cm, and/or with weight values ≤0 kg or ≥45 kg were also excluded from the study. We also excluded data that were considered outliers according to the 2006 WHO Child Growth Standards recommended range values, which were values <-6 or >6 HAZ for stunting, <-5 or >5 WHZ for wasting, and <-6 or >5 WAZ for underweight.
Replication	This is an observational study using many years of survey and report data and in principle could be replicated. Due to the time required to extract, process, and geo-locate all data, as well as to run the statistical models, we have not undertaken an explicit replication analysis.
Randomization	Randomization was not relevant to this study. This analysis is an observational mapping study and there were no experimental groups.
Blinding	Blinding was not relevant to this study, as it was an observational study using survey and report data.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Mapping disparities in education across low- and middle-income countries

https://doi.org/10.1038/s41586-019-1872-1

Received: 9 November 2018

Accepted: 12 November 2019

Published online: 25 December 2019

Open access

Local Burden of Disease Educational Attainment Collaborators*

Educational attainment is an important social determinant of maternal, newborn, and child health[1–3]. As a tool for promoting gender equity, it has gained increasing traction in popular media, international aid strategies, and global agenda-setting[4–6]. The global health agenda is increasingly focused on evidence of precision public health, which illustrates the subnational distribution of disease and illness[7,8]; however, an agenda focused on future equity must integrate comparable evidence on the distribution of social determinants of health[9–11]. Here we expand on the available precision SDG evidence by estimating the subnational distribution of educational attainment, including the proportions of individuals who have completed key levels of schooling, across all low- and middle-income countries from 2000 to 2017. Previous analyses have focused on geographical disparities in average attainment across Africa or for specific countries, but—to our knowledge—no analysis has examined the subnational proportions of individuals who completed specific levels of education across all low- and middle-income countries[12–14]. By geolocating subnational data for more than 184 million person-years across 528 data sources, we precisely identify inequalities across geography as well as within populations.

Education, as a social determinant of health, is closely linked to several facets of the Sustainable Development Goals (SDGs) of the United Nations[2]. In addition to the explicit focus of SDG 4 on educational attainment, improved gender equality (SDG 5) and maternal, newborn, and child health (SDG 3) have well-documented associations with increased schooling[15–17]. In 2016, after years of deprioritization, aid to education reached its highest level since 2002[18]. Despite this shift, only 22% of aid to basic education—defined as primary and lower-secondary—went to low-income countries in 2016 compared to 36% in 2002[19]. This reflects a persistent pattern in which the distribution of aid does not align with the greatest need, even at the national level. Beyond international aid, domestic policy is also a crucial tool for expanding access to education, especially at higher levels. However, policy-makers often do not have access to a rigorous evidence base at a subnational level. This analysis presents the subnational distribution of education to support the growing evidence base of precision public health data, which shows widespread disparity of health outcomes as well as their social determinants.

Mapping education across gender

Despite widespread improvement in educational attainment since 2000, gender disparity persists in 2017 in many regions. Figure 1 illustrates the mean number of years of education and the proportion of individuals with no primary school attainment for men and women of reproductive age (15–49 years) in 2017. The average educational attainment is very low across much of the Sahel region of sub-Saharan Africa, consistent with previously published data[14]. In 2017, there was a large gender disparity in many regions, with men attaining higher average education across central and western sub-Saharan Africa and South Asia. Considerable variation remains between the highest- and lowest-performing administrative units within countries in 2017. For Uganda in 2017, this indicator ranged from 1.9 years of education (95% uncertainty interval, 0.8–3.0 years) in rural Kotido to 11.1 years (10.1–12 years) in Kampala, the capital city. Figure 1b, d displays the proportion of men and women aged 15–49 years who have not completed primary school. By considering the variation within populations in different locations, these maps help to identify areas with large populations in the vulnerable lower end of the attainment distribution. We estimated large improvements in the proportions of individuals who have completed primary school in Mexico and China. However, across much of the world women in this age group failed to complete primary school at a much higher rate than their male counterparts.

Despite continued lack of gender parity in education among the reproductive age group, vast progress towards parity has been made among the 20–24 age group. Extended Data Fig. 2 further examines gender parity in 2000 and 2017. This figure highlights two additional advantages of our analytic framework. First, we examined a younger group aged 20–24 years. Although education in this group is less directly relevant to maternal, newborn, and child health than education in the full window of reproductive age, these estimates allowed us to capture how the landscape of education has shifted over time (that is, across successive cohorts) and is therefore more likely to pick up improvements to access and retention in education systems that have been made since 2000. Second, we illustrate the probability that this estimated ratio is credibly different from 1 (parity between sexes) given the full uncertainty in our data and model. In 2000, we estimated that men completed schooling at a higher rate than women across much

*A list of participants and their affiliations appears in the online version of the paper.

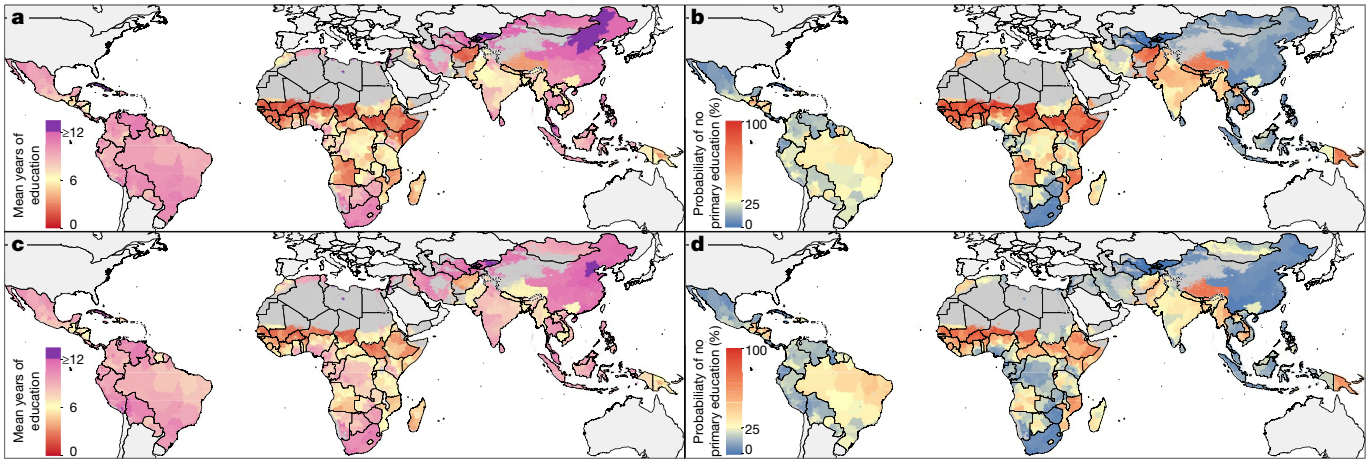


Fig. 1 | Average educational attainment and proportion of individuals with no completed primary education at the first administrative level and absolute difference between women and men aged 15–49 years. a–d, Mean educational attainment for women (**a**) and men (**c**) and the proportion of individuals with no primary school education for women (**b**) and men (**d**) aged 15–49 years in 2017. Maps were produced using ArcGIS Desktop 10.6.

of the world, particularly for primary school education (that is, the probability that the parity ratio is greater than 1 was over 95%). This was true in most countries for both primary and secondary completion rates, but especially so in Burundi, Angola, Uganda, and Afghanistan (Extended Data Fig. 2a, c). By 2017, many countries moved significantly towards parity in both secondary and primary completion rates with the exception of large regions within central and western sub-Saharan Africa (Extended Data Fig. 2b, d).

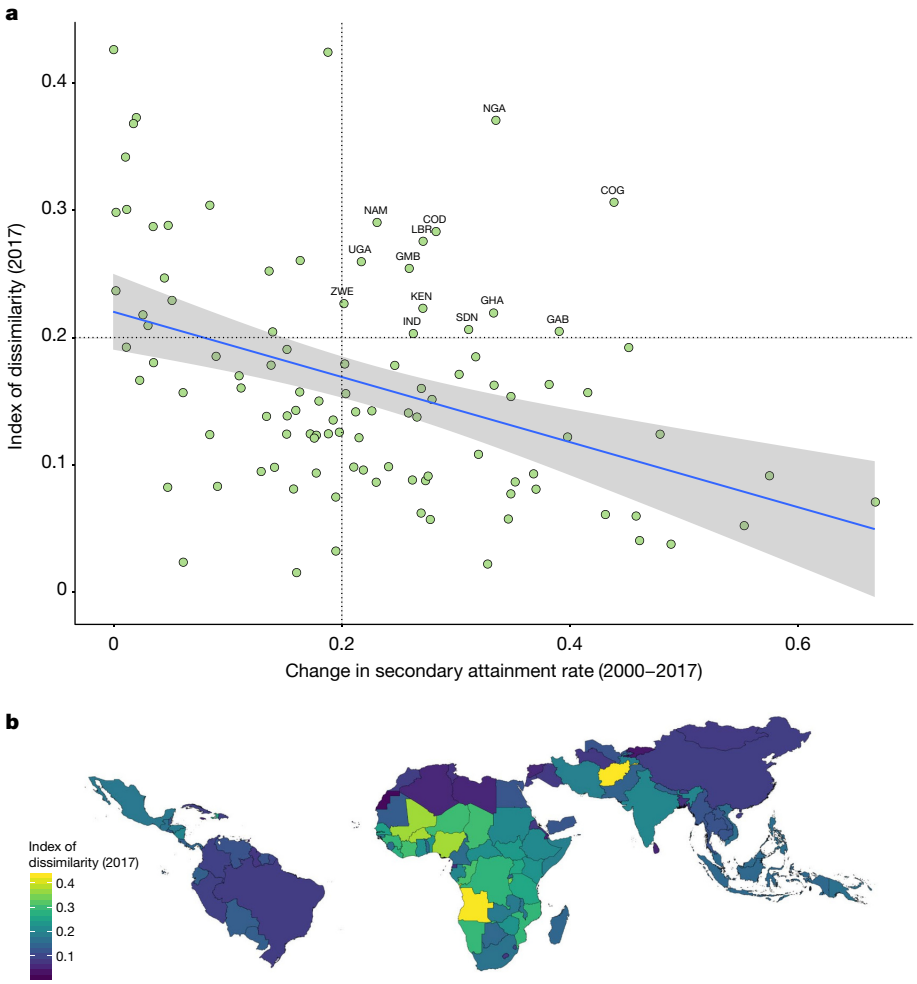


Fig. 2 | National progress in secondary attainment rates for women aged 20–24 years compared with the national index of dissimilarity in 2017. a, Change in secondary attainment rates for women age 20–24 years between 2000 and 2017 compared with the national index of dissimilarity in 2017 (simple linear regression lines are included). **b**, Map of the national index of dissimilarity in 2017. Maps were produced using ArcGIS Desktop 10.6.

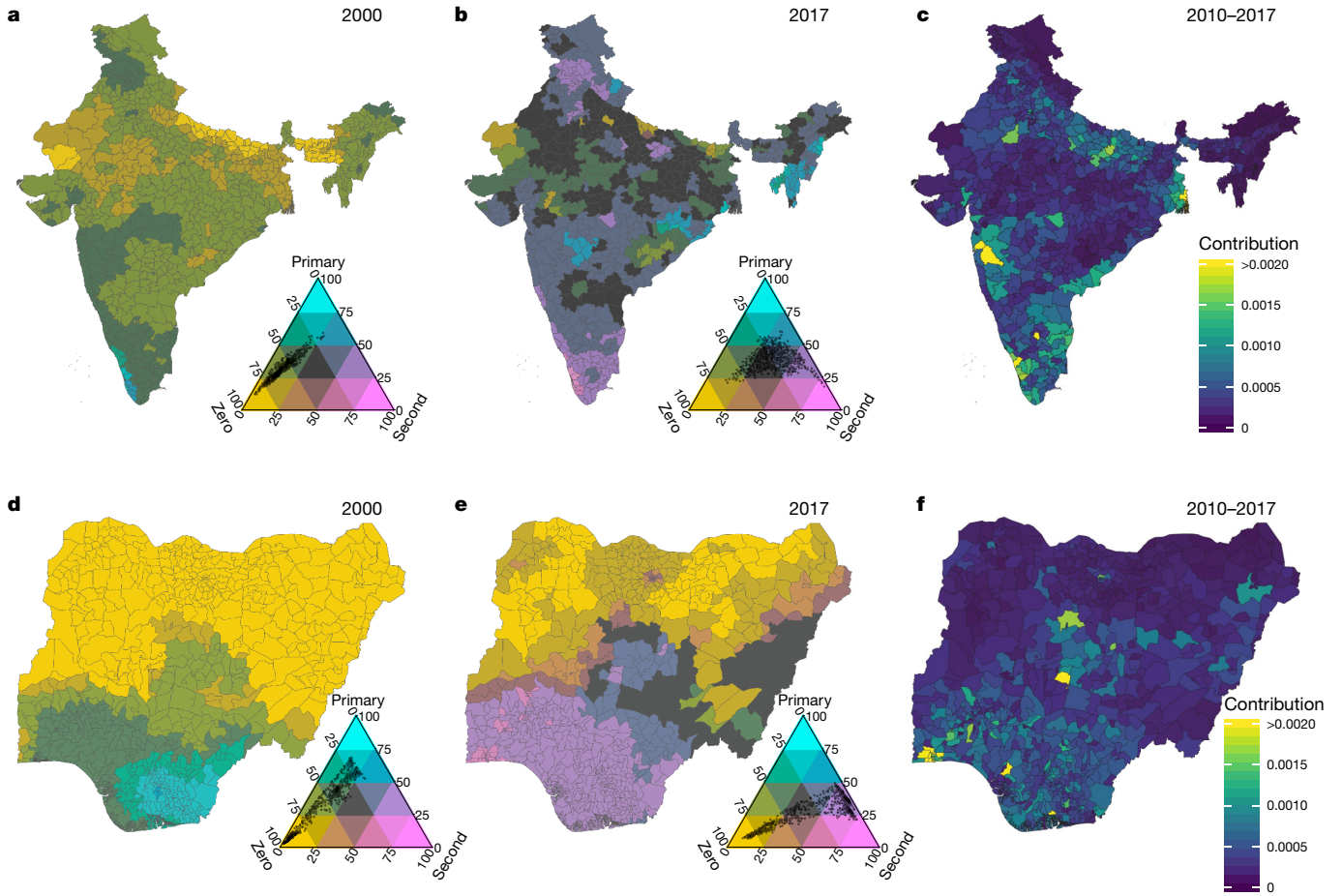


Fig. 3 | Attainment rates and contributions to national change in secondary rates for women aged 20–24 years in India and Nigeria, 2000–2017. **a**, **b**, Attainment rates for women aged 20–24 years in 2000 (**a**) and 2017 (**b**) at the second administrative level in India. **c**, Additive contributions of changes in the attainment rates at the second administrative level to change in the rate at the national level between 2000 and 2017 in India. **d**, **e**, Attainment rates for women aged 20–24 years in 2000 and 2017 at the second administrative level in Nigeria. **f**, Additive contributions of changes in the attainment rates at the second administrative level to change in the rate at the national level between 2000 and 2017 in Nigeria. On all ternary maps, the 'Zero' category includes all individuals with either no schooling or some primary schooling without completion. Maps were produced using ArcGIS Desktop 10.6.

Inequalities within and between countries

The subnational estimates of attainment presented here enable a closer examination of within-country inequality and associated trends over time. Figure 2 plots the national change in secondary attainment rates for women aged 20–24 years with the index of dissimilarity across second administrative-level units in 2017. The index of dissimilarity is an intuitive measure of geographical inequality that can be interpreted as the percentage of women with secondary attainment that would have to move in order to equalize secondary rates across all subnational districts. We estimated that countries that experienced more national progress over the period tended to be more spatially equal in 2017. However, the top-right quadrant of the graph highlights several countries that experienced substantial national progress yet remain some of the most geographically unequal countries today.

We further examined national progress between 2000 and 2017 in two such countries, India and Nigeria, where rates of secondary attainment increased from 10.9% (8.5–12.5%) to 37.2% (33.6–41.1%) and from 11.5% (6.2–18.3%) to 45.0% (37.0–52.5%), respectively (Fig. 3). The geographical distribution between two cohorts—women aged 20–24 years in 2000 and 2017—was analysed by examining all proportions simultaneously (Fig. 3a, b). We estimate that there has been a massive shift towards primary and secondary completion coupled with greater geographical variability in completion rates (that is, spread of the dots that represent subnational units in the legend). The majority of the 2017 cohort living in the northwest and northeast of India never completed secondary school. Urban centres in the south, such as Bangalore and Mumbai, have seen considerable progress compared with more rural regions. In Nigeria, we estimate substantial national improvement; however, the country remained one of the most spatially unequal in 2017 (Fig. 3d, e). The more-urban south, particularly around Lagos, experienced much faster progress than the more-rural north. The implications of the population distribution were explored by decomposing the improvement in the national rate of secondary completion since 2000 for each country into the additive contributions of rate changes at the second administrative level (Fig. 3c, f). This demonstrates that national progress was largely driven by improvements in populous urban regions (particularly Maharashtra, India, and Lagos, Nigeria), underscoring the importance of how subnational progress (or lack thereof) contributes differentially to narratives surrounding national change.

Discussion and limitations

We have built on previous modelling efforts that focused on the geographical distribution of average education[14] by extending our estimation to the distribution of attainment, highlighting not only average attainment but also the proportions of individuals who completed key levels of schooling that are central to policy efforts. As we demonstrate, throughout much of the world women lag behind their male

counterparts, and there is significant heterogeneity across subnational regions. Countries such as South Africa, Peru, and Colombia have seen tremendous improvement since 2000 in the proportion of the young adult population who have completed secondary school. As this trend continues, it will be important to focus not only on attainment but also on quality of education. However, many young women across the world still faced obstacles to attaining even a basic level of education in 2017 (Extended Data Fig. 3). This represents a missed opportunity for the global health community to focus on a well-studied determinant of maternal, newborn, and child health. Even with only marginal returns to health in the short term, studies suggest that, on average, communities will also see increased human capital, social mobility, and less engagement in child marriage or early childbearing[20,21].

Children and adolescents do not complete formal schooling for many reasons. Many factors differentially affect girls, such as cost, late or no school enrolment, forced withdrawal of married adolescents, and the social influence of family members concerning the traditional roles of girls and women[4,20,22,23]. A critical step is acknowledging that commercialization in the area of education typically leads to higher inequity[24]. Treating public education as a societal good by increasing access, particularly in underserved rural communities, reduces inequality. Identifying areas that are stagnating or worsening, particularly in the realm of basic education for young women across the world, is an important first step to targeted, long-term reform efforts that will ultimately have widespread benefits for equity in health and development.

Many recent international calls to improve the social determinants of health have stated that measurement of inequity within countries is critical to understanding and tracking the problem, noting that geography is an increasingly important dimension of inequity[24–26]. Where people are born greatly determines their life chances, and continuing to consider development and human capital formation on a national level is insufficient[24]. The goal of this analysis is to identify local areas that may have experienced negligible improvements, but further rigorous research is required to contextualize these patterns within the unique mix of structural obstacles that each community faces. There are many indirect costs for attending school and each disadvantaged area that we identify in our analysis may experience them in different ways. These include the demand for children to work, the opportunity or monetary costs of attending school, distance to school, lack of compulsory education requirements, high fees for attendance, political instability, and many other forces. Overcoming these obstacles to improve educational attainment alone will not necessarily result in a more-educated and healthy population for each country as highly educated individuals may be more likely to emigrate, resulting in 'brain drain'. This is especially true for countries that have been economically crippled over the past two decades and may lack the economic capacity to absorb a more highly educated labour force. Opening access to education will need to be coupled with economic reforms, both internationally and domestically, if countries are to fully experience dividends in human capital and health.

Over the next decade of the SDG agenda, it will be important to maintain the progress that has been made to reprioritise investment in education systems. There remains an alarming lack of distributional accountability in aid, especially to basic education, for which most funding is not going to the countries that need it most[19]. Connections between educational attainment and health offer promising opportunities for co-financing initiatives. For example, USAID recently invested US$90 million in HIV funding to the construction of secondary schools in sub-Saharan Africa. Global health leaders have noted the need to invest in precise data systems and eliminate data gaps to effectively target resources, develop equitable policy, and track accountability[7]. Our analysis provides a robust evidence base for such decision-making and advocacy. Decades of research on the effect of basic education on maternal, newborn, and child health positions this issue squarely in the purview of the global health agenda. It is crucial for the global health community to invest in long-term, sustainable improvement in the underlying distribution of human capital, as this is the only way to truly influence health equity across generations.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1872-1.

1. UNESCO. Meeting our commitments to gender equality in education. *Global Education Monitoring Report*. https://unesdoc.unesco.org/ark:/48223/pf0000261593 (2018).
2. United Nations. *Transforming our World: the 2030 Agenda for Sustainable Development* (UN, 2015).
3. Lim, S. S. et al. Measuring human capital: a systematic analysis of 195 countries and territories, 1990–2016. *Lancet* **392**, 1217–1234 (2018).
4. Yousafzai, M. & Lamb, C. *I am Malala: The Girl Who Stood Up for Education and Was Shot by The Taliban* (Weidenfeld & Nicolson, 2013).
5. Gates, M. *The Moment of Lift: How Empowering Women Changes The World* (Flatiron Books, 2019).
6. United Nations. *Youth and the 2030 Agenda for Sustainable Development* (UN, 2018).
7. Annan, K. Data can help to end malnutrition across Africa. *Nature* **555**, 7 (2018).
8. Horton, R. Offline: in defence of precision public health. *Lancet* **392**, 1504 (2018).
9. Dowell, S. F., Blazes, D. & Desmond-Hellmann, S. Four steps to precision public health. *Nature* **540**, 189–191 (2016).
10. Osgood-Zimmerman, A. et al. Mapping child growth failure in Africa between 2000 and 2015. *Nature* **555**, 41–47 (2018).
11. Golding, N. et al. Mapping under-5 and neonatal mortality in Africa, 2000–15: a baseline analysis for the Sustainable Development Goals. *Lancet* **390**, 2171–2182 (2017).
12. Bosco, C. et al. Exploring the high-resolution mapping of gender-disaggregated development indicators. *J. R. Soc. Interface* **14**, 20160825 (2017).
13. Roberts, D. A. et al. Benchmarking health system performance across regions in Uganda: a systematic analysis of levels and trends in key maternal and child health interventions, 1990–2011. *BMC Med.* **13**, 285 (2015).
14. Graetz, N. et al. Mapping local variation in educational attainment across Africa. *Nature* **555**, 48–53 (2018).
15. Caldwell, J. C. How is greater maternal education translated into lower child mortality? *Health Transit. Rev.* **4**, 224–229 (1994).
16. Caldwell, J. C. Education as a factor in mortality decline: an examination of Nigerian data. *Popul. Stud.* **33**, 395–413 (1979).
17. Gakidou, E., Cowling, K., Lozano, R. & Murray, C. J. Increased educational attainment and its effect on child mortality in 175 countries between 1970 and 2009: a systematic analysis. *Lancet* **376**, 959–974 (2010).
18. UNESCO. *UNESCO Operational Definition Of Basic Education. Thematic Framework* (UNESCO, 2007).
19. UNESCO. *Aid to Education: A Return to Growth?* (UNESCO, 2018).
20. LeVine, R. A., LeVine, S., Schnell-Anzola, B., Rowe, M. L. & Dexter, E. *Literacy and Mothering: How Women's Schooling Changes the Lives of the World's Children* (Oxford Univ. Press, 2012).
21. Abel, G. J., Barakat, B., Kc, S. & Lutz, W. Meeting the Sustainable Development Goals leads to lower world population growth. *Proc. Natl Acad. Sci. USA* **113**, 14294–14299 (2016).
22. UNESCO. *Reducing Global Poverty Through Universal Primary and Secondary Education. Out-Of-School Children, Adolescents And Youth: Global Status And Trends Policy Paper 32/Fact Sheet* **44** (2017).
23. Jejeebhoy, S. J. *Women's Education, Autonomy, and Reproductive Behaviour: Experience From Developing Countries* (Clarendon, 1995).
24. Marmot, M., Friel, S., Bell, R., Houweling, T. A. & Taylor, S. Closing the gap in a generation: health equity through action on the social determinants of health. *Lancet* **372**, 1661–1669 (2008).
25. Kim, J. Y. *World Bank Group President Jim Yong Kim Speech at the 2017 Annual Meetings Plenary*. http://www.worldbank.org/en/news/speech/2017/10/13/wbg-president-jim-yong-kim-speech-2017-annual-meetings-plenary-session (13 October 2017).
26. UNESCO. *Is real progress being made in the equitable provision of education? #PISAresults* http://www.iiep.unesco.org/en/real-progress-being-made-equitable-provision-education-pisaresults-3915 (UNESCO, 2017).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this license, visit http://creativecommons.org/licenses/by/4.0/.

© The Author(s) 2019

Local Burden of Disease Educational Attainment Collaborators

Nicholas Graetz[1,358], Lauren Woyczynski[1,358], Katherine F. Wilson[1], Jason B. Hall[1], Kalkidan Hassen Abate[2], Foad Abd-Allah[3], Oladimeji M. Adebayo[4], Victor Adekanmbi[5], Mahdi Afshari[6], Olufemi Ajumobi[7,8], Tomi Akinyemiju[9,10], Fares Alahdab[11], Ziyad Al-Aly[12,13], Jacqueline Elizabeth Alcalde Rabanal[14], Mehran Alijanzadeh[15], Vahid Alipour[16], Khalid Altirkawi[17], Mohammadreza Amiresmaili[18], Nahla Hamed Anber[19], Catalina Liliana Andrei[20], Mina Anjomshoa[21], Carl Abelardo T. Antonio[22,23], Jalal Arabloo[16], Olatunde Aremu[24], Krishna K. Aryal[25], Mehran Asadi-Aliabadi[26], Suleman Atique[27], Marcel Ausloos[28], Ashish Awasthi[29], Beatriz Paulina Ayala Quintanilla[30,31], Samad Azari[16], Alaa Badawi[32,33], Joseph Adel Mattar Banoub[34], Suzanne Lyn Barker-Collo[35], Anthony Barnett[36], Neeraj Bedi[37,38], Derrick A. Bennett[39], Natalia V. Bhattacharjee[1], Krittika Bhattacharyya[40,41], Suraj Bhattarai[42], Zulfiqar A. Bhutta[43,44], Ali Bijani[45], Boris Bikbov[46], Gabrielle Britton[47], Roy Burstein[1], Zahid A. Butt[48,49], Rosario Cárdenas[50], Félix Carvalho[51,52], Carlos A. Castañeda-Orjuela[53,54], Franz Castro[55], Ester Cerin[36,56], Jung-Chen Chang[57], Michael L. Collison[1], Cyrus Cooper[58,59], Michael A. Cork[1], Farah Daoud[1], Rajat Das Gupta[60,61], Nicole Davis Weaver[1], Jan-Walter De Neve[62], Kebede Deribe[63,64], Beruk Berhanu Desalegn[65], Aniruddha Deshpande[1], Melaku Desta[66,67], Meghnath Dhimal[68], Daniel Diaz[67,69], Mesfin Tadese Dinberu[70], Shirin Djalalinia[71], Manisha Dubey[72], Eleonora Dubljanin[73], Andre R. Durães[74,75], Laura Dwyer-Lindgren[1,76], Lucas Earl[1], Mohammad Ebrahimi Kalan[77], Ziad El-Khatib[78,79], Babak Eshrati[80,81], Mahbobeh Faramarzi[82], Mohammad Fareed[83], Andre Faro[84], Seyed-Mohammad Fereshtehnejad[85,86], Eduarda Fernandes[87], Irina Filip[88,89], Florian Fischer[90], Takeshi Fukumoto[91,92], Jose A. García[93], Paramjit Singh Gill[94], Tiffany K. Gill[95], Philimon N. Gona[96], Sameer Vali Gopalani[97,98], Ayman Grada[99], Yuming Guo[100,101], Rajeev Gupta[102,103], Vipin Gupta[104], Arvin Haj-Mirzaian[105,106], Arya Haj-Mirzaian[105,107], Randah R. Hamadeh[108], Samer Hamidi[109], Mehedi Hasan[61], Hamid Yimam Hassen[110,111], Delia Hendrie[112], Andualem Henok[110], Nathaniel J. Henry[1], Bernardo Hernández Prado[1,76], Claudiu Herteliu[28], Michael K. Hole[113], Naznin Hossain[114,115], Mehdi Hosseinzadeh[116,117], Guoqing Hu[118], Olayinka Stephen Ilesanmi[119], Seyed Sina Naghibi Irvani[120], Sheikh Mohammed Shariful Islam[121,122], Neda Izadi[123], Mihajlo Jakovljevic[124], Ravi Prakash Jha[125], John S. Ji[126,127], Jost B. Jonas[128,129], Zahra Jorjoran Shushtari[130], Jacek Jerzy Jozwiak[131], Tanuj Kanchan[132], Amir Kasaeian[133,134], Ali Kazemi Karyani[135], Peter Njenga Keiyoro[136], Chandrasekharan Nair Kesavachandran[137], Yousef Saleh Khader[138], Morteza Abdullatif Khafaie[139], Ejaz Ahmad Khan[140], Mona M. Khater[141], Aliasghar A. Kiadaliri[142], Daniel N. Kiirithio[143], Yun Jin Kim[144], Ruth W. Kimokoti[145], Damaris K. Kinyoki[1,76], Adnan Kisa[146], Soewarta Kosen[147], Ai Koyanagi[148,149], Kewal Krishan[150], Barthelemy Kuate Defo[151,152], Manasi Kumar[153,154], Pushpendra Kumar[155], Faris Hasan Lami[156], Paul H. Lee[157], Aubrey J. Levine[1], Shanshan Li[100], Yu Liao[158,159], Lee-Ling Lim[160,161], Stefan Listl[162,163], Jaifred Christian F. Lopez[159,164], Marek Majdan[165], Reza Majdzadeh[166,167], Azeem Majeed[168], Reza Malekzadeh[169,170], Mohammad Ali Mansournia[171], Francisco Rogerlândio Martins-Melo[172], Anthony Masaka[173], Benjamin Ballard Massenburg[174], Benjamin K. Mayala[1], Kala M. Mehta[175], Walter Mendoza[176], George A. Mensah[177,178], Tuomo J. Meretoja[179,180], Tomislav Mestrovic[181,182], Ted R. Miller[112,183], G. K. Mini[184,185], Erkin M. Mirrakhimov[186,187], Babak Moazen[62,188], Dara K. Mohammad[189,190], Aso Mohammad Darwesh[191], Shafiu Mohammed[62,192], Farnam Mohebi[193,194], Ali H. Mokdad[1,76], Lorenzo Monasta[195], Yoshan Moodley[196], Mahmood Moosazadeh[197], Ghobad Moradi[198,199], Maziar Moradi-Lakeh[26], Paula Moraga[200], Lidia Morawska[201], Shane Douglas Morrison[202], Jonathan F. Mosser[1], Seyyed Meysam Mousavi[203,204], Christopher J. L. Murray[1,76], Ghulam Mustafa[205,206], Azin Nahvijou[207], Farid Najafi[208], Vinay Nangia[209], Duduzile Edith Ndwandwe[210], Ionut Negoi[211,212], Ruxandra Irina Negoi[213,214], Josephine W. Ngunjiri[215], Cuong Tat Nguyen[216], Long Hoang Nguyen[217], Dina Nur Anggraini Ningrum[218,219], Jean Jacques Noubiap[178], Malihe Nourollahpour Shiadeh[220], Peter S. Nyasulu[221], Felix Akpojene Ogbo[222], Andrew T. Olagunju[223,224], Bolajoko Olubukunola Olusanya[225], Jacob Olusegun Olusanya[225], Obinna E. Onwujekwe[226], Doris D. V. Ortega-Altamirano[14], Eduardo Ortiz-Panozo[227,228], Simon Øverland[229,230], Mahesh P. A.[231], Adrian Pana[28,232], Songhomitra Panda-Jonas[233], Sanghamitra Pati[234], George C. Patton[235,236], Norberto Perico[237], David M. Pigott[1,76], Meghdad Pirsaheb[135], Maarten J. Postma[238,239], Akram Pourshams[169], Swayam Prakash[240], Parul Puri[241], Mostafa Qorbani[242], Amir Radfar[243,244], Fakher Rahim[245,246], Vafa Rahimi-Movaghar[247], Mohammad Hifz Ur Rahman[248], Fatemeh Rajati[135], Chhabi Lal Ranabhat[249,250], David Laith Rawaf[251,252], Salman Rawaf[168,253], Robert C. Reiner Jr[1,76], Giuseppe Remuzzi[237], Andre M. N. Renzaho[254,255], Satar Rezaei[135], Aziz Rezapour[16], Carlos Rios-González[256,257], Leonardo Roever[258], Luca Ronfani[195], Gholamreza Roshandel[169,259], Ali Rostami[260], Enrico Rubagotti[261,262], Nafis Sadat[1], Ehsan Sadeghi[135], Yahya Safari[263], Rajesh Sagar[264], Nasir Salam[265], Payman Salamati[247], Yahya Salimi[208,266], Hamideh Salimzadeh[169], Abdallah M. Samy[267], Juan Sanabria[268,269], Milena M. Santric Milicevic[270,271], Benn Sartorius[76,272], Brijesh Sathian[273,274], Arundhati R. Sawant[275,276], Lauren E. Schaeffer[1], Megan F. Schipp[1], David C. Schwebel[277], Anbissa Muleta Senbeta[278], Sadaf G. Sepanlou[169,170], Masood Ali Shaikh[279], Mehran Shams-Beyranvand[280,281], Morteza Shamsizadeh[282], Kiomars Sharafi[135], Rajesh Sharma[283], Jun She[284], Aziz Sheikh[285,286], Mika Shigematsu[287], Soraya Siabani[288,289], Dayane Gabriele Alves Silveira[290,291], Jasvinder A. Singh[292,293], Dhirendra Narain Sinha[294,295], Vegard Skirbekk[296], Amber Sligar[1], Badr Hasan Sobaih[297,298], Moslem Soofi[266], Joan B. Soriano[299,300], Ireneous N. Soyiri[301,302], Chandrashekhar T. Sreeramareddy[303], Agus Sudaryanto[304,305], Mu'awiyyah Babale Sufiyan[306], Ipsita Sutradhar[61], PN Sylaja[307,308], Rafael Tabarés-Seisdedos[309,310], Birkneh Tilahun Tadesse[311,312], Mohamad-Hani Temsah[297,313], Abdullah Sulieman Terkawi[314,315], Belay Tessema[316], Zemenu Tadesse Tessema[317], Kavumpurathu Raman Thankappan[318], Roman Topor-Madry[319,320], Marcos Roberto Tovani-Palone[321], Bach Xuan Tran[322], Lorainne Tudor Car[323], Irfan Ullah[324,325], Olalekan A. Uthman[326], Pascual R. Valdez[327,328], Yousef Veisani[329], Francesco S. Violante[330,331], Vasily Vlassov[332], Sebastian Vollmer[333,334], Giang Thu Vu[217], Yasir Waheed[335], Yuan-Pang Wang[336], John C. Wilkinson[1], Andrea Sylvia Winkler[337,338], Charles D. A. Wolfe[339,340], Tomohide Yamada[341], Alex Yeshaneh[342], Paul Yip[343,344], Engida Yisma[345], Naohiro Yonemoto[346], Mustafa Z. Younis[347,348], Mahmoud Yousefifard[349], Chuanhua Yu[350,351], Sojib Bin Zaman[352,353], Jianrong Zhang[354], Yunquan Zhang[355,356], Sanjay Zodpey[29], Emmanuela Gakidou[1,76,357] & Simon I. Hay[1,76,357]*

[1]Institute for Health Metrics and Evaluation, University of Washington, Seattle, WA, USA. [2]Department of Population and Family Health, Jimma University, Jimma, Ethiopia. [3]Department of Neurology, Cairo University, Cairo, Egypt. [4]Department of Medicine, University College Hospital, Ibadan, Nigeria. [5]School of Medicine, Cardiff University, Cardiff, UK. [6]Department of Community Medicine, Zabol University of Medical Sciences, Zabol, Iran. [7]School of Community Health Sciences, University of Nevada, Reno, NV, USA. [8]National Malaria Elimination Program, Federal Ministry of Health, Abuja, Nigeria. [9]Duke Global Health Institute, Duke University, Durham, NC, USA. [10]Department of Population Health Sciences, Duke University, Durham, NC, USA. [11]Evidence Based Practice Center, Mayo Clinic Foundation for Medical Education and Research, Rochester, MN, USA. [12]Internal Medicine Department, Washington University in St Louis, St Louis, MO, USA. [13]Clinical Epidemiology Center, VA Saint Louis Health Care System, Department of Veterans Affairs, St Louis, MO, USA. [14]Center for Health Systems Research, National Institute of Public Health, Cuernavaca, Mexico. [15]Qazvin University of Medical Sciences, Qazvin, Iran. [16]Health Management and Economics Research Center, Iran University of Medical Sciences, Tehran, Iran. [17]King Saud University, Riyadh, Saudi Arabia. [18]Department of Health Management, Policy and Economics, Kerman University of Medical Sciences, Kerman, Iran. [19]Faculty of Medicine, Mansoura University, Mansoura, Egypt. [20]Carol Davila University of Medicine and Pharmacy, Bucharest, Romania. [21]Social Determinants of Health Research Center, Rafsanjan University of Medical Sciences, Rafsanjan, Iran. [22]Department of Health Policy and Administration, University of the Philippines Manila, Manila, The Philippines. [23]Department of Applied Social Sciences, Hong Kong Polytechnic University, Hong Kong, China. [24]School of Health Sciences, Birmingham City University, Birmingham, UK. [25]Monitoring Evaluation and Operational Research Project, ABT Associates Nepal, Lalitpur, Nepal. [26]Preventive Medicine and Public Health Research Center, Iran University of Medical Sciences, Tehran, Iran. [27]Department of Health Informatics, University of Ha'il, Ha'il, Saudi Arabia. [28]Department of Statistics and Econometrics, Bucharest University of Economic Studies, Bucharest, Romania. [29]Indian Institute of Public Health, Public Health Foundation of India, Gurugram, India. [30]The Judith Lumley Centre, La Trobe University, Melbourne, Victoria, Australia. [31]General Office for Research and Technological Transfer, Peruvian National Institute of Health, Lima, Peru. [32]Public Health Risk Sciences Division, Public Health Agency of Canada, Toronto, Ontario, Canada. [33]Department of Nutritional Sciences, University of Toronto, Toronto, Ontario, Canada. [34]Faculty of Medicine, Alexandria University, Alexandria, Egypt. [35]School of Psychology, University of Auckland, Auckland, New Zealand. [36]Mary MacKillop Institute for Health Research, Australian Catholic University, Melbourne, Victoria, Australia. [37]Department of Community Medicine, Gandhi Medical College Bhopal, Bhopal, India. [38]Jazan University, Jazan, Saudi Arabia. [39]Nuffield Department of Population Health, University of Oxford, Oxford, UK. [40]Department of Statistical and Computational Genomics, National Institute of Biomedical Genomics, Kalyani, India. [41]Department of Statistics, University of Calcutta, Kolkata, India. [42]Department of Global Health, Global Institute for Interdisciplinary Studies, Kathmandu, Nepal. [43]Centre for Global Child Health, University of Toronto, Toronto, Ontario, Canada. [44]Centre of Excellence in Women and Child Health, Aga Khan University, Karachi, Pakistan. [45]Social Determinants of Health Research Center, Babol University of Medical Sciences, Babol, Iran. [46]Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Ranica, Italy. [47]Center for Neuroscience, Instituto de Investigaciones Científicas y Servicios de Alta Tecnología (INDICASAT AIP), Panama, Panama. [48]School of Public Health and Health Systems, University of Waterloo, Waterloo, Ontario, Canada. [49]Al Shifa School of Public Health, Al Shifa Trust Eye Hospital, Rawalpindi, Pakistan. [50]Department of Population and Health, Metropolitan Autonomous University, Mexico City, Mexico. [51]Institute of Public Health, University of Porto, Porto, Portugal. [52]Applied Molecular Biosciences Unit, University of Porto, Porto, Portugal. [53]Colombian National Health Observatory, National Institute of Health, Bogota, Colombia. [54]Epidemiology and Public Health Evaluation Group, National University of Colombia, Bogota, Colombia. [55]Gorgas Memorial Institute for Health Studies, Panama, Panama. [56]School of Public Health, University of Hong Kong, Hong Kong, China. [57]College of Medicine, National Taiwan University, Taipei, Taiwan. [58]Medical Research Council Lifecourse Epidemiology Unit, University of Southampton, Southampton, UK. [59]Department of Rheumatology, University of Oxford, Oxford, UK. [60]Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, USA. [61]James P. Grant School of Public Health, Brac University, Dhaka, Bangladesh. [62]Heidelberg Institute of Global Health, Heidelberg University, Heidelberg, Germany. [63]Department of Global Health and Infection, Brighton and Sussex Medical School, Brighton, UK. [64]School of Public Health, Addis Ababa University, Addis, Ababa, Ethiopia. [65]School of Nutrition, Food Science and Technology, Hawassa University, Hawassa, Ethiopia. [66]Department of Midwifery, Debre Markos University, Debre, Markos, Ethiopia. [67]Faculty of Veterinary Medicine and Zootechnics, Autonomous University of Sinaloa, Culiacan Rosales, Mexico. [68]Health Research Section, Nepal Health Research Council, Kathmandu, Nepal. [69]Center of Complexity Sciences, National Autonomous University of Mexico, Mexico City, Mexico. [70]Department of Midwifery, Debre Berhan University, Debre Berhan, Ethiopia. [71]Deputy of Research and Technology, Ministry of Health and Medical Education, Tehran, Iran. [72]United Nations World Food Programme, New Delhi, India. [73]Faculty of Medicine, University of Belgrade, Belgrade, Serbia. [74]Department of Internal Medicine, Bahia School of Medicine and Public Health, Salvador, Brazil. [75]Medical

Board, Roberto Santos General Hospital, Salvador, Brazil. [76]Department of Health Metrics Sciences, School of Medicine, University of Washington, Seattle, WA, USA. [77]Epidemiology Department, Florida International University, Miami, FL, USA. [78]Department of Public Health Sciences, Karolinska Institutet, Stockholm, Sweden. [79]World Health Programme, Université du Québec en Abitibi-Témiscamingue, Rouyn-Noranda, Quebec, Canada. [80]Center of Communicable Disease Control, Ministry of Health and Medical Education, Tehran, Iran. [81]School of Public Health, Arak University of Medical Sciences, Arak, Iran. [82]Babol University of Medical Sciences, Babol, Iran. [83]College of Medicine, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. [84]Department of Psychology, Federal University of Sergipe, Sao Cristovao, Brazil. [85]Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden. [86]Division of Neurology, University of Ottawa, Ottawa, Ontario, Canada. [87]REQUIMTE/LAQV, University of Porto, Porto, Portugal. [88]Psychiatry Department, Kaiser Permanente, Fontana, CA, USA. [89]Department of Health Sciences, A. T. Still University, Mesa, AZ, USA. [90]Department of Population Medicine and Health Services Research, Bielefeld University, Bielefeld, Germany. [91]Department of Dermatology, Kobe University, Kobe, Japan. [92]Gene Expression & Regulation Program, The Wistar Institute, Philadelphia, PA, USA. [93]Ramón de la Fuente Muñiz National Institute of Psychiatry, Mexico City, Mexico. [94]Unit of Academic Primary Care, University of Warwick, Coventry, UK. [95]Adelaide Medical School, University of Adelaide, Adelaide, South Australia, Australia. [96]Nursing and Health Sciences Department, University of Massachusetts Boston, Boston, MA, USA. [97]Department of Biostatistics and Epidemiology, University of Oklahoma, Oklahoma City, OK, USA. [98]Department of Health and Social Affairs, Government of the Federated States of Micronesia, Palikir, Federated States of Micronesia. [99]School of Medicine, Boston University, Boston, MA, USA. [100]School of Public Health and Preventive Medicine, Monash University, Melbourne, Victoria, Australia. [101]Department of Epidemiology and Biostatistics, Zhengzhou University, Zhengzhou, China. [102]Academics and Research Department, Rajasthan University of Health Sciences, Jaipur, India. [103]Department of Medicine, Mahatma Gandhi University of Medical Sciences & Technology, Jaipur, India. [104]Department of Anthropology, University of Delhi, Delhi, India. [105]Department of Pharmacology, Tehran University of Medical Sciences, Tehran, Iran. [106]Obesity Research Center, Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [107]Department of Radiology, Johns Hopkins University, Baltimore, MD, USA. [108]Department of Family and Community Medicine, Arabian Gulf University, Manama, Bahrain. [109]School of Health and Environmental Studies, Hamdan Bin Mohammed Smart University, Dubai, United Arab Emirates. [110]Department of Public Health, Mizan-Tepi University, Tepi, Ethiopia. [111]Unit of Epidemiology and Social Medicine, University Hospital Antwerp, Antwerp, Belgium. [112]School of Public Health, Curtin University, Perth, Western Australia, Australia. [113]Department of Pediatrics, University of Texas Austin, Austin, TX, USA. [114]Department of Pharmacology and Therapeutics, Dhaka Medical College, Dhaka, Bangladesh. [115]Department of Pharmacology, Bangladesh Industrial Gases Limited, Tangail, Bangladesh. [116]Department of Computer Engineering, Islamic Azad University, Tehran, Iran. [117]Computer Science Department, University of Human Development, Sulaimaniyah, Iraq. [118]Department of Epidemiology and Health Statistics, Central South University, Changsha, China. [119]Department of Community Medicine, University of Ibadan, Ibadan, Nigeria. [120]Research Institute for Endocrine Sciences, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [121]Institute for Physical Activity and Nutrition, Deakin University, Burwood, Victoria, Australia. [122]Sydney Medical School, University of Sydney, Sydney, New South Wales, Australia. [123]Department of Epidemiology, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [124]Department of Health Care and Public Health, Sechenov First Moscow State Medical University, Moscow, Russia. [125]Department of Community Medicine, Banaras Hindu University, Varanasi, India. [126]Environmental Research Center, Duke Kunshan University, Kunshan, China. [127]Nicholas School of the Environment, Duke University, Durham, NC, USA. [128]Department of Ophthalmology, Heidelberg University, Heidelberg, Germany. [129]Beijing Institute of Ophthalmology, Beijing Tongren Hospital, Beijing, China. [130]Social Determinants of Health Research Center, University of Social Welfare and Rehabilitation Sciences, Tehran, Iran. [131]Department of Family Medicine and Public Health, University of Opole, Opole, Poland. [132]Department of Forensic Medicine and Toxicology, All India Institute of Medical Sciences, Jodhpur, India. [133]Hematology-Oncology and Stem Cell Transplantation Research Center, Tehran University of Medical Sciences, Tehran, Iran. [134]Pars Advanced and Minimally Invasive Medical Manners Research Center, Iran University of Medical Sciences, Tehran, Iran. [135]Research Center for Environmental Determinants of Health, Kermanshah University of Medical Sciences, Kermanshah, Iran. [136]ODeL Campus, University of Nairobi, Nairobi, Kenya. [137]CSIR-Indian Institute of Toxicology Research, Council of Scientific & Industrial Research, Lucknow, India. [138]Department of Public Health, Jordan University of Science and Technology, Irbid, Jordan. [139]Social Determinants of Health Research Center, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. [140]Epidemiology and Biostatistics Department, Health Services Academy, Islamabad, Pakistan. [141]Department of Medical Parasitology, Cairo University, Cairo, Egypt. [142]Clinical Epidemiology Unit, Lund University, Lund, Sweden. [143]Research and Data Solutions, Synotech Consultants, Nairobi, Kenya. [144]School of Medicine, Xiamen University Malaysia, Sepang, Malaysia. [145]Department of Nutrition, Simmons University, Boston, MA, USA. [146]School of Health Sciences, Kristiania University College, Oslo, Norway. [147]Independent Consultant, Jakarta, Indonesia. [148]CIBERSAM, San Juan de Dios Sanitary Park, Sant Boi De Llobregat, Spain. [149]Catalan Institution for Research and Advanced Studies (ICREA), Barcelona, Spain. [150]Department of Anthropology, Panjab University, Chandigarh, India. [151]Department of Social and Preventive Medicine, University of Montreal, Montreal, Quebec, Canada. [152]Department of Demography, University of Montreal, Montreal, Quebec, Canada. [153]Department of Psychiatry, University of Nairobi, Nairobi, Kenya. [154]Division of Psychology and Language Sciences, University College London, London, UK. [155]International Institute for Population Sciences, Mumbai, India. [156]Department of Community and Family Medicine, University of Baghdad, Baghdad, Iraq. [157]School of Nursing, Hong Kong Polytechnic University, Hong Kong, China. [158]Department of Medical Statistics and Epidemiology, Sun Yat-sen University, Guangzhou, China. [159]Alliance for Improving Health Outcomes Inc, Quezon City, The Philippines. [160]Department of Medicine, University of Malaya, Kuala Lumpur, Malaysia. [161]Department of Medicine and Therapeutics, The Chinese University of Hong Kong, Shatin, China. [162]Department of Dentistry, Radboud University, Nijmegen, The Netherlands. [163]Section for Translational Health Economics, Heidelberg University Hospital, Heidelberg, Germany. [164]Department of Epidemiology and Biostatistics, University of the Philippines Manila, Manila, The Philippines. [165]Department of Public Health, Trnava University, Trnava, Slovakia. [166]Community-Based Participatory-Research Center (CBPR), Tehran University of Medical Sciences, Tehran, Iran. [167]Knowledge Utilization Research Center (KURC), Tehran University of Medical Sciences, Tehran, Iran. [168]Department of Primary Care and Public Health, Imperial College London, London, UK. [169]Digestive Diseases Research Institute, Tehran University of Medical Sciences, Tehran, Iran. [170]Non-communicable Diseases Research Center, Shiraz University of Medical Sciences, Shiraz, Iran. [171]Department of Epidemiology and Biostatistics, Tehran University of Medical Sciences, Tehran, Iran. [172]Campus Caucaia, Federal Institute of Education, Science and Technology of Ceará, Caucaia, Brazil. [173]Public Health Department, Botho University-Botswana, Gaborone, Botswana. [174]Division of Plastic Surgery, University of Washington, Seattle, WA, USA. [175]Department of Epidemiology and Biostatistics, University of California San Francisco, San Francisco, CA, USA. [176]Peru Country Office, United Nations Population Fund (UNFPA), Lima, Peru. [177]Center for Translation Research and Implementation Science, National Institutes of Health, Bethesda, MD, USA. [178]Department of Medicine, University of Cape Town, Cape Town, South Africa. [179]Breast Surgery Unit, Helsinki University Hospital, Helsinki, Finland. [180]University of Helsinki, Helsinki, Finland. [181]Clinical Microbiology and Parasitology Unit, Dr Zora Profozic Polyclinic, Zagreb, Croatia. [182]University Centre Varazdin, University North, Varazdin, Croatia. [183]Pacific Institute for Research & Evaluation, Calverton, MD, USA. [184]Achutha Menon Centre for Health Science Studies, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India. [185]Global Institute of Public Health (GIPH), Ananthapuri Hospitals and Research Centre, Trivandrum, India. [186]Faculty of Internal Medicine, Kyrgyz State Medical Academy, Bishkek, Kyrgyzstan. [187]Department of Atherosclerosis and Coronary Heart Disease, National Center of Cardiology and Internal Disease, Bishkek, Kyrgyzstan. [188]Institute of Addiction Research (ISFF), Frankfurt University of Applied Sciences, Frankfurt, Germany. [189]Department of Food Technology, College of Agriculture, Salahaddin University-Erbil, Erbil, Iraq. [190]Department of Medicine Huddinge, Karolinska Institutet, Stockholm, Sweden. [191]Department of Information Technology, University of Human Development, Sulaimaniyah, Iraq. [192]Health Systems and Policy Research Unit, Ahmadu Bello University, Zaria, Nigeria. [193]Non-communicable Diseases Research Center, Tehran University of Medical Sciences, Tehran, Iran. [194]Iran National Institute of Health Research, Tehran University of Medical Sciences, Tehran, Iran. [195]Clinical Epidemiology and Public Health Research Unit, Burlo Garofolo Institute for Maternal and Child Health, Trieste, Italy. [196]Department of Public Health Medicine, University of Kwazulu-Natal, Durban, South Africa. [197]Health Sciences Research Center, Mazandaran University of Medical Sciences, Sari, Iran. [198]Social Determinants of Health Research Center, Kurdistan University of Medical Sciences, Sanandaj, Iran. [199]Department of Epidemiology and Biostatistics, Kurdistan University of Medical Sciences, Sanandaj, Iran. [200]Department of Mathematical Sciences, University of Bath, Bath, UK. [201]International Laboratory for Air Quality and Health, Queensland University of Technology, Brisbane, Queensland, Australia. [202]Department of Surgery, University of Washington, Seattle, WA, USA. [203]Department of Health Management and Economics, Tehran University of Medical Sciences, Tehran, Iran. [204]Health Management Research Center, Baqiyatallah University of Medical Sciences, Tehran, Iran. [205]Department of Pediatric Medicine, Nishtar Medical University, Multan, Pakistan. [206]Department of Pediatrics & Pediatric Pulmonology, Institute of Mother & Child Care, Multan, Pakistan. [207]Cancer Research Center, Tehran University of Medical Sciences, Tehran, Iran. [208]Department of Epidemiology & Biostatistics, Kermanshah University of Medical Sciences, Kermanshah, Iran. [209]Suraj Eye Institute, Nagpur, India. [210]Cochrane South Africa, South African Medical Research Council, Cape Town, South Africa. [211]General Surgery, Carol Davila University of Medicine and Pharmacy Bucharest, Bucharest, Romania. [212]General Surgery, Emergency Hospital of Bucharest, Bucharest, Romania. [213]Anatomy and Embryology, Carol Davila University of Medicine and Pharmacy, Bucharest, Romania. [214]Cardiology, Cardio-Aid, Bucharest, Romania. [215]Department of Biological Sciences, University of Embu, Embu, Kenya. [216]Institute for Global Health Innovations, Duy Tan University, Hanoi, Vietnam. [217]Center of Excellence in Behavioral Medicine, Nguyen Tat Thanh University, Ho Chi Minh City, Vietnam. [218]Public Health Department, Universitas Negeri Semarang, Kota Semarang, Indonesia. [219]Graduate Institute of Biomedical Informatics, Taipei Medical University, Taipei City, Taiwan. [220]Mazandaran University of Medical Sciences, Sari, Iran. [221]Faculty of Medicine & Health Sciences, Stellenbosch University, Cape Town, South Africa. [222]UCIBIO, University of Porto, Porto, Portugal. [223]Department of Psychiatry and Behavioural Neurosciences, McMaster University, Hamilton, Ontario, Canada. [224]Department of Psychiatry, University of Lagos, Lagos, Nigeria. [225]Centre for Healthy Start Initiative, Lagos, Nigeria. [226]Department of Pharmacology and Therapeutics, University of Nigeria Nsukka, Enugu, Nigeria. [227]Center for Population Health Research, National Institute of Public Health, Cuernavaca, Mexico. [228]School of Health and Welfare, Jönköping University, Jönköping, Sweden. [229]Division of Mental and Physical

Health, Norwegian Institute of Public Health, Bergen, Norway. [230]Department of Psychosocial Science, University of Bergen, Bergen, Norway. [231]Department of Respiratory Medicine, Jagadguru Sri Shivarathreeswara Academy of Health Education and Research, Mysore, India. [232]Health Outcomes, Center for Health Outcomes & Evaluation, Bucharest, Romania. [233]Augenpraxis Jonas, Heidelberg University, Heidelberg, Germany. [234]Regional Medical Research Centre, Indian Council of Medical Research, Bhubaneswar, India. [235]Department of Paediatrics, University of Melbourne, Melbourne, Victoria, Australia. [236]Population Health, Murdoch Children's Research Institute, Melbourne, Victoria, Australia. [237]Istituto di Ricerche Farmacologiche Mario Negri IRCCS, Bergamo, Italy. [238]Department of Economics and Business, University of Groningen, Groningen, The Netherlands. [239]University Medical Center Groningen, University of Groningen, Groningen, The Netherlands. [240]Department of Nephrology, Sanjay Gandhi Postgraduate Institute of Medical Sciences, Lucknow, India. [241]Population Studies, International Institute for Population Sciences, Mumbai, India. [242]Non-communicable Diseases Research Center, Alborz University of Medical Sciences, Karaj, Iran. [243]College of Medicine, University of Central Florida, Orlando, FL, USA. [244]College of Graduate Health Sciences, A. T. Still University, Mesa, AZ, USA. [245]Thalassemia and Hemoglobinopathy Research Center, Ahvaz Jundishapur University of Medical Sciences, Ahvaz, Iran. [246]Metabolomics and Genomics Research Center, Tehran University of Medical Sciences, Tehran, Iran. [247]Sina Trauma and Surgery Research Center, Tehran University of Medical Sciences, Tehran, Iran. [248]Department of Public Health and Mortality Studies, International Institute for Population Sciences, Mumbai, India. [249]Policy Research Institute, Kathmandu, Nepal. [250]Institute for Poverty Alleviation and International Development, Yonsei University, Wonju, South Korea. [251]WHO Collaborating Centre for Public Health Education and Training, Imperial College London, London, UK. [252]University College London Hospitals, London, UK. [253]Academic Public Health, Public Health England, London, UK. [254]Translational Health Research Institute, Western Sydney University, Penrith, New South Wales, Australia. [255]School of Social Sciences and Psychology, Western Sydney University, Penrith, New South Wales, Australia. [256]Research Directorate, Nihon Gakko University, Fernando De La Mora, Paraguay. [257]Research Direction, Universidad Nacional de Caaguazú, Coronel Oviedo, Paraguay. [258]Department of Clinical Research, Federal University of Uberlândia, Uberlândia, Brazil. [259]Golestan Research Center of Gastroenterology and Hepatology, Golestan University of Medical Sciences, Gorgan, Iran. [260]Infectious Diseases and Tropical Medicine Research Center, Babol University of Medical Sciences, Babol, Iran. [261]Centro de Investigación Palmira, Agrosavia, Palmira, Colombia. [262]Department of Ocean Science and Engineering, Southern University of Science and Technology, Shenzhen, China. [263]Kermanshah University of Medical Sciences, Kermanshah, Iran. [264]Department of Psychiatry, All India Institute of Medical Sciences, New Delhi, India. [265]Department of Pathology, Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia. [266]Social Development and Health Promotion Research Center, Kermanshah University of Medical Sciences, Kermanshah, Iran. [267]Department of Entomology, Ain Shams University, Cairo, Egypt. [268]Department of Surgery, Marshall University, Huntington, WV, USA. [269]Department of Nutrition and Preventive Medicine, Case Western Reserve University, Cleveland, OH, USA. [270]Institute of Social Medicine, University of Belgrade, Belgrade, Serbia. [271]Centre-School of Public Health and Health Management, University of Belgrade, Belgrade, Serbia. [272]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, London, UK. [273]Surgery Department, Hamad Medical Corporation, Doha, Qatar. [274]Faculty of Health & Social Sciences, Bournemouth University, Bournemouth, UK. [275]University of Alabama at Birmingham, Birmingham, AL, USA. [276]Dr D. Y. Patil University, Pune, India. [277]Department of Psychology, University of Alabama at Birmingham, Birmingham, AL, USA. [278]Department of Food Science and Nutrition, Jigjiga University, Jigjiga, Ethiopia. [279]Independent Consultant, Karachi, Pakistan. [280]School of Medicine, Dezful University of Medical Sciences, Dezful, Iran. [281]School of Medicine, Alborz University of Medical Sciences, Karaj, Iran. [282]Chronic Diseases (Home Care) Research Center, Hamadan University of Medical Sciences, Hamadan, Iran. [283]University School of Management and Entrepreneurship, Delhi Technological University, New Delhi, India. [284]Department of Pulmonary Medicine, Fudan University, Shanghai, China. [285]Centre for Medical Informatics, University of Edinburgh, Edinburgh, UK. [286]Division of General Internal Medicine, Harvard University, Boston, MA, USA. [287]National Institute of Infectious Diseases, Tokyo, Japan. [288]Department of Health Education & Promotion, Kermanshah University of Medical Sciences, Kermanshah, Iran. [289]School of Health, University of Technology Sydney, Sydney, New South Wales, Australia. [290]Brasília University, Brasília, Brazil. [291]Department of the Health Industrial Complex and Innovation in Health, Federal Ministry of Health, Brasília, Brazil. [292]Department of Epidemiology, University of Alabama at Birmingham, Birmingham, AL, USA. [293]Department of Medicine, University of Alabama at Birmingham, Birmingham, AL, USA. [294]Department of Epidemiology, School of Preventive Oncology, Patna, India. [295]Department of Epidemiology, Healis Sekhsaria Institute for Public Health, Mumbai, India. [296]Centre for Fertility and Health, Norwegian Institute of Public Health, Bergen, Norway. [297]Department of Pediatrics, King Saud University, Riyadh, Saudi Arabia. [298]Pediatric Department, King Khalid University Hospital, Riyadh, Saudi Arabia. [299]Hospital Universitario de la Princesa, Autonomous University of Madrid, Madrid, Spain. [300]Centro de Investigación Biomédica en Red Enfermedades Respiratorias (CIBERES), Madrid, Spain. [301]Usher Institute of Population Health Sciences and Informatics, University of Edinburgh, Edinburgh, UK. [302]Hull York Medical School, University of Hull, Hull, UK. [303]Division of Community Medicine, International Medical University, Kuala Lumpur, Malaysia. [304]Department of Nursing, Muhammadiyah University of Surakarta, Kartasura, Indonesia. [305]Department of Public Health, China Medical University, Taichung, Taiwan. [306]Department of Community Medicine, Ahmadu Bello University, Zaria, Nigeria. [307]Neurology Department, Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India. [308]Sree Chitra Tirunal Institute for Medical Sciences and Technology, Trivandrum, India. [309]Department of Medicine, University of Valencia, Valencia, Spain. [310]Carlos III Health Institute, Biomedical Research Networking Center for Mental Health Network (CIBERSAM), Madrid, Spain. [311]Department of Pediatrics, Hawassa University, Hawassa, Ethiopia. [312]International Vaccine Institute, Seoul, South Korea. [313]College of Medicine, Alfaisal University, Riyadh, Saudi Arabia. [314]Department of Anesthesiology, Perioperative, and Pain Medicine, University of Virginia, Charlottesville, VA, USA. [315]Department of Anesthesiology, King Farah Medical City, Riyadh, Saudi Arabia. [316]Department of Medical Microbiology, University of Gondar, Gondar, Ethiopia. [317]Department of Epidemiology and Biostatistics, University of Gondar, Gondar, Ethiopia. [318]Department of Public Health and Community Medicine, Central University of Kerala, Kasaragod, India. [319]Faculty of Health Sciences, Jagiellonian University Medical College, Krakow, Poland. [320]The Agency for Health Technology Assessment and Tariff System, Warsaw, Poland. [321]Department of Pathology and Legal Medicine, University of São Paulo, Ribeirão Preto, Brazil. [322]Department of Health Economics, Hanoi Medical University, Hanoi, Vietnam. [323]Lee Kong Chian School of Medicine, Nanyang Technological University, Singapore, Singapore. [324]Gomal Center of Biochemistry and Biotechnology, Gomal University, Dera Ismail Khan, Pakistan. [325]TB Culture Laboratory, Mufti Mehmood Memorial Teaching Hospital, Dera Ismail Khan, Pakistan. [326]Division of Health Sciences, University of Warwick, Coventry, UK. [327]Argentine Society of Medicine, Ciudad de Buenos Aires, Argentina. [328]Velez Sarsfield Hospital, Buenos Aires, Argentina. [329]Psychosocial Injuries Research Center, Ilam University of Medical Sciences, Ilam, Iran. [330]Department of Medical and Surgical Sciences, University of Bologna, Bologna, Italy. [331]Occupational Health Unit, Sant'Orsola Malpighi Hospital, Bologna, Italy. [332]Department of Health Care Administration and Economics, National Research University Higher School of Economics, Moscow, Russia. [333]Department of Global Health and Population, Harvard University, Boston, MA, USA. [334]Department of Economics, University of Göttingen, Göttingen, Germany. [335]Foundation University Medical College, Foundation University Islamabad, Islamabad, Pakistan. [336]Department of Psychiatry, University of São Paulo, São Paulo, Brazil. [337]Institute of Health and Society, University of Oslo, Oslo, Norway. [338]Department of Neurology, Technical University of Munich, Munich, Germany. [339]School of Population Health & Environmental Sciences, King's College London, London, UK. [340]NIHR Biomedical Research Centre, Guy's and St Thomas' Hospital and Kings College London, London, UK. [341]Department of Diabetes and Metabolic Diseases, University of Tokyo, Tokyo, Japan. [342]Wolkite University, Wolkite, Ethiopia. [343]Centre for Suicide Research and Prevention, University of Hong Kong, Hong Kong, China. [344]Department of Social Work and Social Administration, University of Hong Kong, Hong Kong, China. [345]School of Allied Health Sciences, Addis Ababa University, Addis Ababa, Ethiopia. [346]Department of Psychopharmacology, National Center of Neurology and Psychiatry, Tokyo, Japan. [347]Health Economics & Finance, Global Health, Jackson State University, Jackson, MS, USA. [348]School of Medicine, Tsinghua University, Peking, China. [349]Prevention of Cardiovascular Disease Research Center, Shahid Beheshti University of Medical Sciences, Tehran, Iran. [350]Global Health Institute, Wuhan University, Wuhan, China. [351]Department of Epidemiology and Biostatistics, Wuhan University, Wuhan, China. [352]Department of Medicine, Monash University, Melbourne, Victoria, Australia. [353]Maternal and Child Health Division, International Centre for Diarrhoeal Disease Research, Bangladesh, Dhaka, Bangladesh. [354]George Warren Brown School, Washington University in St Louis, St Louis, MO, USA. [355]School of Public Health, Wuhan University of Science and Technology, Wuhan, China. [356]Hubei Province Key Laboratory of Occupational Hazard Identification and Control, Wuhan University of Science and Technology, Wuhan, China. [357]These authors contributed equally: Nicholas Graetz, Lauren Woyczynski. [358]These authors jointly supervised this work: Emmanuela Gakidou, Simon I. Hay. *e-mail: sihay@uw.edu

Methods

Overview

Using a Bayesian model-based geostatistical framework and synthesizing geolocated data from 528 household and census datasets, this analysis provides subnational estimates of mean numbers years of education and the proportion of the population who attained key levels of education for women of reproductive age (15–49 years), women aged 20–24 years, and equivalent male age bins between 2000 and 2017 in 105 countries across all low- and middle-income countries (LMICs). Countries were selected for inclusion in this analysis using the socio-demographic index (SDI) published in the Global Burden of Disease (GBD) study[27]. The SDI is a measure of development that combines education, fertility, and poverty. Countries in the middle, lower-middle, or low SDI quintiles were included, with several exceptions. Albania, Bosnia, and Moldova were excluded despite middle SDI status due to geographical discontinuity with other included countries and lack of available survey data. Libya, Malaysia, Panama, and Turkmenistan were included despite higher-middle SDI status to create better geographical continuity. We did not analyse American Samoa, Federated States of Micronesia, Fiji, Kiribati, Marshall Islands, Samoa, Solomon Islands, or Tonga, where no available survey data could be sourced. Analytical steps are described below, and additional details can be found in the Supplementary Information.

Data

We compiled a database of survey and census datasets that contained geocoding of subnational administrative boundaries or GPS coordinates for sampled clusters. These included datasets from 528 sources (see Supplementary Table 2). These sources comprised at least one data source for all but two countries on our list of LMICs: Western Sahara and French Guiana. We chose to exclude these two countries from our analysis; 42 of 105 included countries have only subnational administrative level data. We extracted demographic, education, and sample design variables. The coding of educational attainment varies across survey families. In some surveys, the precise number of years of attainment is not provided, with attainment instead aggregated into categories such as 'primary completion' or 'secondary completion'. In such cases, individuals who report 'primary completion' may have gone on to complete some portion of secondary education, but these additional years of education are not captured in the underlying dataset. Previous efforts to examine trends in mean years of education have either assumed that no additional years of education were completed (that is, primary education only) or have used the midpoint between primary and secondary education as a proxy[28]. Trends in the single-year data, however, demonstrate that such assumptions introduce bias in the estimation of attainment trends over time and space, as differences in actual drop-out patterns or binning schema can lead to biased mean estimates[29].

For this analysis, we used a recently developed method that selects a training subset of similar surveys across time and space to estimate the unobserved single-year distribution of binned datasets[29]. In comprehensive tests of cross-validation that leveraged data for which the single-year distributions are observed, this algorithmic approach significantly reduces bias in summary statistics estimated from datasets with binned coding schemes compared to alternatives such as the standard-duration method[28]. The years in all coding schemes were mapped to the country- and year-specific references in the UNESCO International Standard Classification of Education (ISCED) for comparability[30]. We used a top coding of 18 years on all data; this is a common threshold in many surveys that have a cap and it is reasonable to assume that the importance of education for health outcomes (and other related SDGs) greatly diminishes after what is the equivalent of 2 to 3 years of graduate education in most systems.

Data were aggregated to mean years of education attained and the proportions achieving key levels of education. The levels chosen were proportion with zero years, proportion with less than primary school (1–5 years of education), proportion with at least primary school (6–11 years of education), and proportion achieving secondary school or higher (12 or more years of education). A subset of the data for a smaller age bin (20–24 years) was also examined to more closely track temporal shifts. Equivalent age bins were aggregated for both women and men to examine disparities in mean years of attainment by sex. Where GPS coordinates were available, data were aggregated to a specific latitude and longitude assuming a simple-random sample, as the cluster is the primary sampling unit for the stratified design survey families, such as the Demographic and Health Survey (DHS) and Multiple Indicator Cluster Survey (MICS). Where only geographical information was available at the level of administrative units, data were aggregated with appropriate weighting according to their sample design. Design effects were estimated using a package for analysing complex survey data in R[31].

Spatial covariates

To leverage strength from locations with observations to the entire spatiotemporal domain, we compiled several 5×5-km^2 raster layers of possible socioeconomic and environmental correlates of education (Supplementary Table 5 and Supplementary Fig. 6). Acquisition of temporally dynamic datasets, where possible, was prioritized to best match our observations and thus predict the changing dynamics of educational attainment. We included nine covariates indexed at the 5×5-km^2 level: access to roads, nighttime lights^{tv}, population^{tv}, growing season, aridity^{tv}, elevation, urbanicity^{tv}, irrigation, and year^{tv} (tv, time-varying covariates). More details, including plots of all covariates, can be found in the Supplementary Information.

Our primary goal is to provide educational attainment predictions across LMICs at a high (local) resolution, and our methods provide the best out-of-sample predictive performance at the expense of inferential understanding. To select covariates and capture possible nonlinear effects and complex interactions between them, an ensemble covariate modelling method was implemented[32]. For each region, three submodels were fitted to our outcomes using all of our covariate data: generalized additive models, boosted regression trees, and lasso regression. Each submodel was fit using fivefold cross-validation to avoid overfitting and the out-of-sample predictions from across the five folds were compiled into a single comprehensive set of predictions from that model. Additionally, the same submodels were also run using 100% of the data and a full set of in-sample predictions were created. The five sets of out-of-sample submodel predictions were fed into the full geostatistical model as predictors when performing the model fit. The in-sample predictions from the submodels were used as the covariates when generating predictions using the fitted full geostatistical model. This methodology maximizes out-of-sample predictive performance at the expense of the ability to provide statistical inference on the relationships between the predictors and the outcome. A recent study has shown that this ensemble approach can improve predictive validity by up to 25% over an individual model[32]. More details on this approach can be found in the Supplementary Information.

The primary goal of using the stacking procedure in our analyses was to maximize the predictive power of the raster covariates by capturing the nonlinear effects and complex interactions between covariates to optimize the model performance. It has previously been suggested[32] that the primary purpose of the submodel predictions is to improve the mean function of the Gaussian process. Although we have determined a way to include the uncertainty from two of our submodels (lasso regression and generalized additive models (GAM)), we have not determined a way to include uncertainty from the boosted regression tree (BRT) submodel into our final estimates. Whereas GAM and lasso regression seek to fit a single model that best describes the relationship

between response variable and some set of predictors, BRT method fits a large number of relatively simple models for which the predictions are then combined to give robust estimates of the response. Although this feature of the BRT model makes it a powerful tool for analysing complex data, quantifying the relative uncertainty contributed by each simple model as well as uncertainty from the complex interactions of the predictor variables is challenging[33,34]. It is worth noting, however, that our out-of-sample validation indicates that the 95% coverage is fairly accurate (for example, closely ranges around 95%) as shown in the figures and table of Supplementary Information section 4.3.2. This indicates that we are not misrepresenting the uncertainty in our final estimates.

Analysis

Geostatistical model. Gaussian and binomial data are modelled within a Bayesian hierarchical modelling framework using a spatially and temporally explicit hierarchical generalized linear regression model to fit the mean number years of education attainment and the proportion of the population who achieved key bins of school in 14 regions across all LMICs as defined in the GBD study (Extended Data Fig. 1). This means we fit 14 independent models for each indicator (for example, the proportion of women with zero years of schooling). GBD study design sought to create regions on the basis of three primary criteria: epidemiological homogeneity, sociodemographic similarity, and geographical contiguity[27]. Fitting our models by these regions has the advantage of allowing for some non-stationarity and non-isotropy in the spatial error term, compared to if we modelled one spatiotemporal random-effect structure over the entire modelling region of all LMICs.

For each Gaussian indicator, we modelled the mean number of years of attainment in each survey cluster, d. Survey clusters are precisely located by their GPS coordinates and year of observation, which we map to a spatial raster location i at time t. We model the mean number of years of attainment as Gaussian data given fixed precision τ and a scaling parameter s_d (defined by the sample size in the observed cluster). As we may have observed multiple data clusters within a given location i at time t, we refer to the mean attainment, μ, within a given cluster d by its indexed location i, and time t as $\mu_{i(d),t(d)}$.

$$\text{edu}_d|\mu_{i(d),t(d)}, s_d, \tau \sim \text{Normal}\left(\mu_{i(d),t(d)}, \tau s_d\right) \ \forall \text{ observed clusters } d$$

$$\mu_{i,t} = \beta_0 + \mathbf{X}_{i,t}\boldsymbol{\beta} + Z_{i,t} + \epsilon_{\text{ctr}(i)} + \epsilon_{i,t} \ \forall \ i \in \text{spatial domain} \ \forall \ t \in \text{time domain}$$

For each binomial indicator, we modelled the number of individuals at a given attainment level in each survey cluster, d. We observed the number of individuals reporting a given attainment level as binomial count data C_d among an observed sample size N_d. As we may have observed multiple data clusters within a given location i at time t, we refer to the probability of attaining that level, p, within a given cluster d by its indexed location i and time t as $p_{i(d),t(d)}$.

$$C_d|p_{i(d),t(d)}, N_d \sim \text{Binomial}\left(p_{i(d),t(d)}, N_d\right) \ \forall \text{ observed clusters } d$$

$$\text{logit}\left(p_{i,t}\right) = \beta_0 + \mathbf{X}_{i,t}\boldsymbol{\beta} + Z_{i,t} + \epsilon_{\text{ctr}(i)} + \epsilon_{i,t} \ \forall \ i \in \text{spatial domain} \ \forall \ t \in \text{time domain}$$

We used a continuation-ratio modelling approach to account for the ordinal data structure of the binomial indicators[35]. To do this, the proportion of the population with zero years of education was modelled using a binomial model. The proportion with less than primary education was modelled as those with less than primary education of those that have more than zero years of education. The same method followed for the proportion of population completing primary education. The proportion achieving secondary school or higher was estimated as the complement of the sum of the three binomial models.

The remaining parameter specification was consistent between all indicators in both binomial and Gaussian models:

$$\sum_{h=1}^{3} \beta_h = 1$$

$$\epsilon_{\text{ctr}} \sim \text{iid Normal}(0, \gamma^2)$$

$$\epsilon_{i,t} \sim \text{iid Normal}(0, \sigma^2)$$

$$\mathbf{Z} \sim \text{GP}(0, \Sigma^{\text{space}} \otimes \Sigma^{\text{time}})$$

$$\Sigma^{\text{space}} = \frac{\omega^2}{\Gamma(\nu)2^{\nu-1}} \times (\kappa D)^{\nu} \times K_{\nu}(\kappa D)$$

$$\Sigma^{\text{time}}_{j,k} = \rho^{|k-j|}$$

For indices d, i, and t, *(index) is the value of * at that index. The probabilities $p_{i,t}$ represent both the annual proportions at the space–time location and the probability that an individual had that level of attainment given that they lived at that particular location. The annual probability $p_{i,t}$ of each indicator (or $\mu_{i,t}$ for the mean indicators) was modelled as a linear combination of the three submodels (GAM, BRT, and lasso regression), rasterized covariate values $X_{i,t}$, a correlated spatiotemporal error term $Z_{i,t}$, country random effects $\epsilon_{\text{ctr}(i)}$ with one unstructured country random effect fit for each country in the modelling region and all sharing a common variance parameter, γ^2, and an independent nugget effect $\epsilon_{i,t}$ with variance parameter σ^2. Coefficients β_h in the three submodels $h = 1, 2, 3$ represent their respective predictive weighting in the mean logit link, while the joint error term $Z_{i,t}$ accounts for residual spatiotemporal autocorrelation between individual data points that remains after accounting for the predictive effect of the submodel covariates, the country-level random effect $\epsilon_{\text{ctr}(i)}$, and the nugget independent error term, $\epsilon_{i,t}$. The purpose of the country-level random effect is to capture spatially unstructured, unobserved country-specific variables, as there are often sharp discontinuities in educational attainment between adjacent countries due to systematic differences in governance, infrastructure, and social policies.

The residuals $Z_{i,t}$ are modelled as a three-dimensional Gaussian process (GP) in space–time centred at zero and with a covariance matrix constructed from a Kronecker product of spatial and temporal covariance kernels. The spatial covariance Σ^{space} is modelled using an isotropic and stationary Matérn function[36], and temporal covariance Σ^{time} as an annual autoregressive (AR1) function over the 18 years represented in the model. In the stationary Matérn function, Γ is the Gamma function, K_{ν} is the modified Bessel function of order $\nu > 0$, $\kappa > 0$ is a scaling parameter, D denotes the Euclidean distance, and ω^2 is the marginal variance. The scaling parameter, κ, is defined to be $\kappa = \sqrt{8\nu}/\delta$ where δ is a range parameter (which is about the distance for which the covariance function approaches 0.1) and ν is a scaling constant, which is set to 2 rather than fit from the data[37,38]. This parameter is difficult to reliably fit, as documented by many other analyses[37,39,40] that set this parameter to 2. The number of rows and the number of columns of the spatial Matérn covariance matrix are equal to the number of spatial mesh points for a given modelling region. In the AR1 function, ρ is the autocorrelation function (ACF), and k and j are points in the time series where $|k-j|$ defines the lag. The number of rows and the number of columns of the AR1 covariance matrix are equal to the number of temporal mesh points (18). The number of rows and the number of columns of the space–time covariance matrix, $\Sigma^{\text{space}} \otimes \Sigma^{\text{time}}$, for a given modelling region are equal to: the number of spatial mesh points × the number of temporal mesh points.

This approach leveraged the residual correlation structure of the data to more accurately predict estimates for locations with no data, while also propagating the dependence in the data through to uncertainty estimates[41]. The posterior distributions were fit using computationally efficient and accurate approximations in R-integrated nested Laplace approximation (INLA) with the stochastic partial differential equations (SPDE) approximation to the Gaussian process residuals using R project version 3.5.1[42–45]. The SPDE approach using INLA has been demonstrated elsewhere, including the estimation of health indicators, particulate air matter, and population age structure[10,11,46,47]. Uncertainty intervals were generated from 1,000 draws (that is, statistically plausible candidate maps)[48] created from the posterior-estimated distributions of modelled parameters. Additional details regarding model and estimation processes can be found in the Supplementary Information.

To transform grid cell-level estimates into a range of information that is useful to a wide constituency of potential users, these estimates were aggregated from the 1,000 candidate maps up to district, provincial, and national levels using 5×5-km^2 population data[49]. This aggregation also enabled the calibration of estimates to national GBD estimates for 2000–2017. This was achieved by calculating the ratio of the posterior mean national-level estimate from each candidate map draw in the analysis to the posterior mean national estimates from GBD, and then multiplying each cell in the posterior sample by this ratio. National-level estimates from this analysis with GBD estimates can be found in Supplementary Table 44.

To illustrate how subnational progress has contributed differentially to national progress (Fig. 3), we decomposed the improvement in the national rate of secondary completion since 2000 for each country into the additive contributions of rate changes at the second administrative level, where C is the national secondary rate change, N is the total number of second-level administrative units, c_i is the population proportion in administrative unit i, and r_i is the rate of secondary attainment in administrative unit i.

$$C = \sum_{i=1}^{N} (c_{i,2017} r_{i,2017}) - (c_{i,2000} r_{i,2000})$$

Although the model can predict at all locations covered by available raster covariates, all final model outputs for which land cover was classified as 'barren or sparsely vegetated' were masked, on the basis of the most recently available Moderate Resolution Imaging Spectroradiometer (MODIS) satellite data (2013), as well as areas in which the total population density was less than 10 individuals per 1×1-km^2 pixel in 2015[50]. This step has led to improved understanding when communicating with data specialists and policy-makers.

Model validation. Models were validated using source-stratified fivefold cross-validation. To offer a more stringent analysis by respecting some of the source and spatial correlation in the data, holdout sets were created by combining sets of data sources (for example, entire survey- or census-years). Model performance was summarized by the bias (mean error), total variance (root-mean-square error) and 95% data coverage within prediction intervals, and the correlation between observed data and predictions. All validation metrics were calculated on the predictions from the fivefold cross-validation. Where possible, estimates from these models were compared against other existing estimates. Furthermore, measures of spatial and temporal autocorrelation pre- and post-modelling were examined to verify correct recognition, fitting, and accounting for the complex spatiotemporal correlation structure in the data. All validation procedures and corresponding results are provided in the Supplementary Information.

Limitations. Our analysis is not without several important limitations. First, almost all data collection tools conflate gender and sex and we therefore do not capture the full distribution of sex or gender separately in our data. We refer throughout to the measurement of 'gender (in) equality', following the usage in SDG 5. Second, it is extremely difficult to quantify quality of education on this scale in a comparable way. Quality is ultimately a large part of the SDG agenda and of utmost importance to achieving equity in opportunity for social mobility. However, many studies across diverse low- and middle-income settings have linked attainment, even very low levels, to measurable improvement in maternal and child health[17]. As our analysis highlights with the proportional indicators, there are still many subnational regions across the world where large proportions do not complete primary school. A third limitation is that we are unable to measure or account for migration. A concept note released from the forthcoming Global Education Monitoring Report 2019 focuses on how migration and displacement affects schooling[51]. Our estimates of the modelled outcome, educational attainment for a particular space–time–age–sex, are demonstrated to be statistically unbiased (Supplementary Information section 4.3); however, interpretation of any change in attainment as a change in the underlying education system could potentially be biased by the effects of migration. It is possible that geographical disparities reflect changes in population composition rather than changes in the underlying infrastructure or education system. Pathways for this change are complex and may be voluntary. Those who manage to receive an education in a low-attainment area may have an increased ability to migrate and choose to do so. This change may also be involuntary, particularly in politically unstable areas where displacement may make geographical changes over time difficult to estimate. A shifting population composition is a general limitation of many longitudinal ecological analyses, but the spatially granular nature of the analyses used here may be more sensitive to the effects of mobile populations.

Our analysis is purely predictive but draws heavily in its motivation from a rich history of literature on the role of education in reducing maternal mortality, improving child health, and increasing human capital. Studies have also demonstrated complex relationships between increased education and a myriad of positive health outcomes, such as HIV risk reductions and spillover effects to other household members[52,53]. The vast majority of these studies are associational and recent attempts at causal analyses have provided more-mixed evidence[54–56]. Although causal analyses of education are very difficult and often rely on situational quasi-experiments, associational analyses using the most comprehensive datasets demonstrate consistent support for the connection between education and health[17,57]. Looking towards future analyses, it will be important to study patterns of change in these data and how they overlap with distributions of health. Lastly, our estimates cannot be seen as a replacement for proper data collection systems, especially for tracking contemporaneous change. Our analysis of uncertainty at a high-resolution may be used to inform investment in more robust data systems and collection efforts, especially if the ultimate goal is to measure and track progress in the quality of schooling.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The findings of this study are supported by data that are available in public online repositories, data that are publicly available upon request from the data provider, and data that are not publicly available owing to restrictions by the data provider, which were used under license for the current study, but may be available from the authors upon reasonable request and permission of the data provider. A detailed table of data sources and availability can be found in Supplementary Table 2. Interactive visualization tools are available at https://vizhub.healthdata.org/lbd/education. All maps presented in this study are generated by the authors; no permissions are required for publication. Administrative

boundaries were retrieved from the Global Administrative Unit Layers (GAUL) dataset, implemented by FAO within the CountrySTAT and Agricultural Market Information System (AMIS) projects[58]. Land cover was retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota[50]. Lakes were retrieved from the Global Lakes and Wetlands Database (GLWD), courtesy of the World Wildlife Fund and the Center for Environmental Systems Research, University of Kassel[59,60]. Populations were retrieved from WorldPop[49,61]. All maps were produced using ArcGIS Desktop 10.6.

Code availability

Our study follows the Guidelines for Accurate and Transparent Health Estimates Reporting (GATHER). All code used for these analyses is available online at http://ghdx.healthdata.org/record/ihme-data/lmic-education-geospatial-estimates-2000-2017, and at http://github.com/ihmeuw/lbd/tree/edu-lmic-2019.

27. GBD 2016 Causes of Death Collaborators. Global, regional, and national age-sex specific mortality for 264 causes of death, 1980–2016: a systematic analysis for the Global Burden of Disease Study 2016. *Lancet* **390**, 1151–1210 (2017).
28. Barro, R. J. & Lee, J. W. A new data set of educational attainment in the world, 1950–2010. *J. Dev. Econ*. **104**, 184–198 (2013).
29. Friedman, J., Graetz, N. & Gakidou, E. Improving the estimation of educational attainment: new methods for assessing average years of schooling from binned data. *PLoS ONE* **13**, e0208019 (2018).
30. UNESCO. *ISCED Mappings* (UNESCO, 2016); http://uis.unesco.org/en/isced-mappings
31. Lumley, T. survey: analysis of complex survey samples. R package v.3.36 https://cran.r-project.org/web/packages/survey.pdf (2019).
32. Bhatt, S. et al. Improved prediction accuracy for disease risk mapping using Gaussian process stacked generalization. *J. R. Soc. Interface* **14**, 20170520 (2017).
33. Elith, J., Leathwick, J. R. & Hastie, T. A working guide to boosted regression trees. *J. Anim. Ecol*. **77**, 802–813 (2008).
34. Leathwick, J., Elith, J., Francis, M., Hastie, T. & Taylor, P. Variation in demersal fish species richness in the oceans surrounding New Zealand: an analysis using boosted regression trees. *Mar. Ecol. Prog. Ser*. **321**, 267–281 (2006).
35. Hosmer, D. W., & Lemeshow, S. in *Applied Logistic Regression* 289–305 (Wiley, 2013).
36. Stein, M. L. *Interpolation of Spatial Data* (Springer, 1999).
37. Lindgren, F. & Rue, H. Bayesian spatial modelling with R-INLA. *J. Stat. Softw*. **63**, 1–25 (2015).
38. Lindgren, F., Rue, H. & Lindström, J. An explicit link between Gaussian fields and Gaussian Markov random fields: the stochastic partial differential equation approach. *J. R. Stat. Soc. B* **73**, 423–498 (2011).
39. Rozanov, Y. A. in *Markov Random Fields* 55–102 (Springer, 1982).
40. Whittle, P. On stationary processes in the plane. *Biometrika* **41**, 434–449 (1954).
41. Diggle, P. & Ribeiro, P. J. *Model-based Geostatistics* (Springer, 2007).
42. Rue, H., Martino, S. & Chopin, N. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *J. R. Stat. Soc. B* **71**, 319–392 (2009).
43. Rue, H. et al. *Bayesian Computing with INLA* (2014); http://www.r-inla.org/
44. Blangiardo, M., Cameletti, M., Baio, G. & Rue, H. Spatial and spatio-temporal models with R-INLA. Spat. *Spatiotemporal Epidemiol*. **7**, 39–55 (2013).
45. Krainski, E. T., Lindgren, F., Simpson, D. & Rue, H. *The R-INLA Tutorial on SPDE Models* (2017); https://inla.r-inla-download.org/r-inla.org/tutorials/spde/spde-tutorial.pdf
46. Cameletti, M., Lindgren, F., Simpson, D. & Rue, H. Spatio-temporal modeling of particulate matter concentration through the SPDE approach. *Adv. Stat. Anal*. **97**, 109–131 (2013).
47. Alegana, V. A. et al. Fine resolution mapping of population age-structures for health and development applications. *J. R. Soc. Interface* **12**, 20150073 (2015).
48. Patil, A. P., Gething, P. W., Piel, F. B. & Hay, S. I. Bayesian geostatistics in health cartography: the perspective of malaria. *Trends Parasitol*. **27**, 246–253 (2011).
49. Tatem, A. J. WorldPop, open data for spatial demography. *Sci. Data* **4**, 170004 (2017).
50. Friedl, M. et al. MCD12Q2 v006. MODIS/Terra+Aqua Land Cover Type Yearly L3 Global 500 m SIN Grid. (NASA EOSDIS Land Processes DAAC, 2019); https://doi.org/10.5067/MODIS/MCD12Q1.006
51. Global Education Monitoring Report. *Concept Note for the 2019 Global Education Monitoring Report on Education and Migration* (2017).
52. Behrman, J. A. The effect of increased primary schooling on adult women's HIV status in Malawi and Uganda: universal primary education as a natural experiment. *Soc. Sci. Med*. **127**, 108–115 (2015).
53. De Neve, J.-W., Fink, G., Subramanian, S. V., Moyo, S. & Bor, J. Length of secondary schooling and risk of HIV infection in Botswana: evidence from a natural experiment. *Lancet Glob. Health* **3**, e470–e477 (2015).
54. McCrary, J. & Royer, H. The effect of female education on fertility and infant health: evidence from school entry policies using exact date of birth. *Am. Econ. Rev*. **101**, 158–195 (2011).
55. Karlsson, O., De Neve, J.-W. & Subramanian, S. V. Weakening association of parental education: analysis of child health outcomes in 43 low- and middle-income countries. *Int. J. Epidemiol*. **48**, 83–97 (2019).
56. De Neve, J.-W. & Fink, G. Children's education and parental old age survival — quasi-experimental evidence on the intergenerational effects of human capital investment. *J. Health Econ*. **58**, 76–89 (2018).
57. Pamuk, E. R., Fuchs, R. & Lutz, W. Comparing relative effects of education and economic resources on infant mortality in developing countries. *Popul. Dev. Rev*. **37**, 637–664 (2011).
58. FAO-UN. *The Global Administrative Unit Layers (GAUL)* (2015); http://www.fao.org/geonetwork/srv/en/metadata.show?id=12691
59. World Wildlife Fund. *Global Lakes and Wetlands Database, Level 3* (2004); https://www.worldwildlife.org/pages/global-lakes-and-wetlands-database
60. Lehner, B. & Döll, P. Development and validation of a global database of lakes, reservoirs and wetlands. *J. Hydrol*. **296**, 1–22 (2004).
61. World Pop. *Data Types* (accessed: 7th July 2017); https://www.worldpop.org/project/list
62. Channan, S., Collins, K. & Emanuel, W. Global mosaics of the standard MODIS land cover type data. University of Maryland and the Pacific Northwest National Laboratory, College Park, Maryland, USA. (2014)

Acknowledgements This work was primarily supported by grant OPP1132415 from the Bill & Melinda Gates Foundation. N.G. is the recipient of a training grant from the National Institute of Child Health and Human Development (T32 HD-007242-36A1).

Author contributions S.I.H. and N.G. conceived and planned the study. K.W. and J.H. extracted, processed, and geo-positioned the data. L.W. and N.G. carried out the statistical analyses. All authors provided intellectual inputs into aspects of this study. N.G., L.W., J.H., and L.E. prepared figures and tables. N.G. wrote the manuscript with assistance by S.B.M., and all authors contributed to subsequent revisions.

Competing interests The authors declare no competing interests.

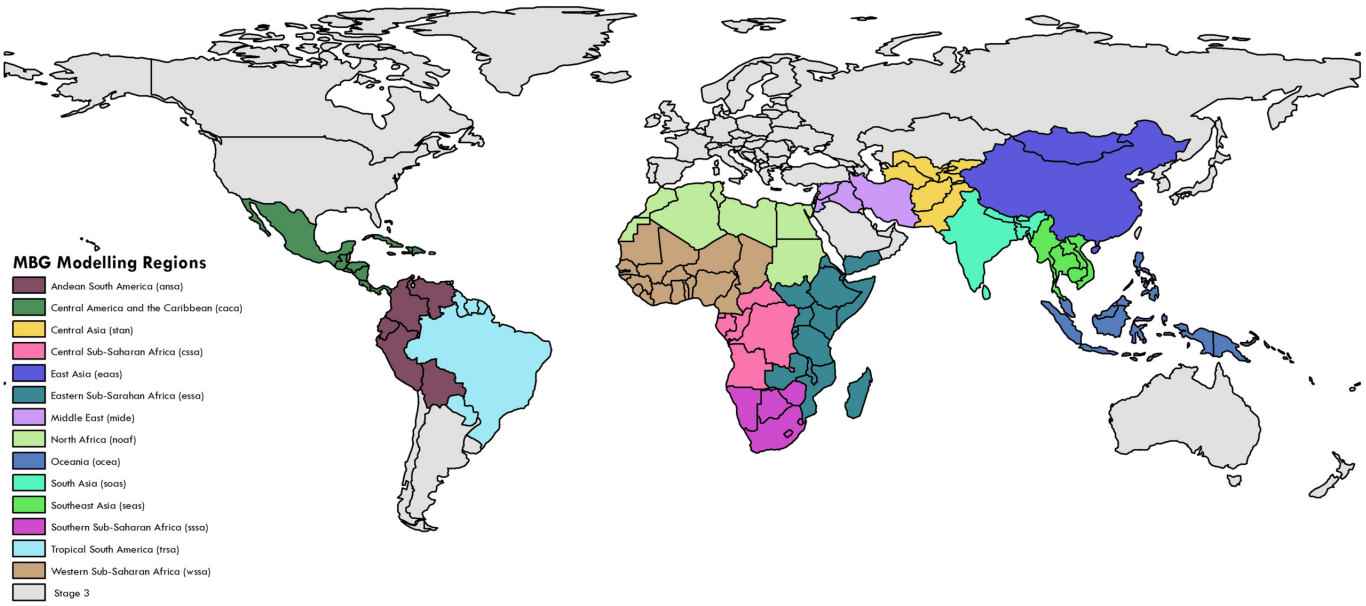
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1872-1.

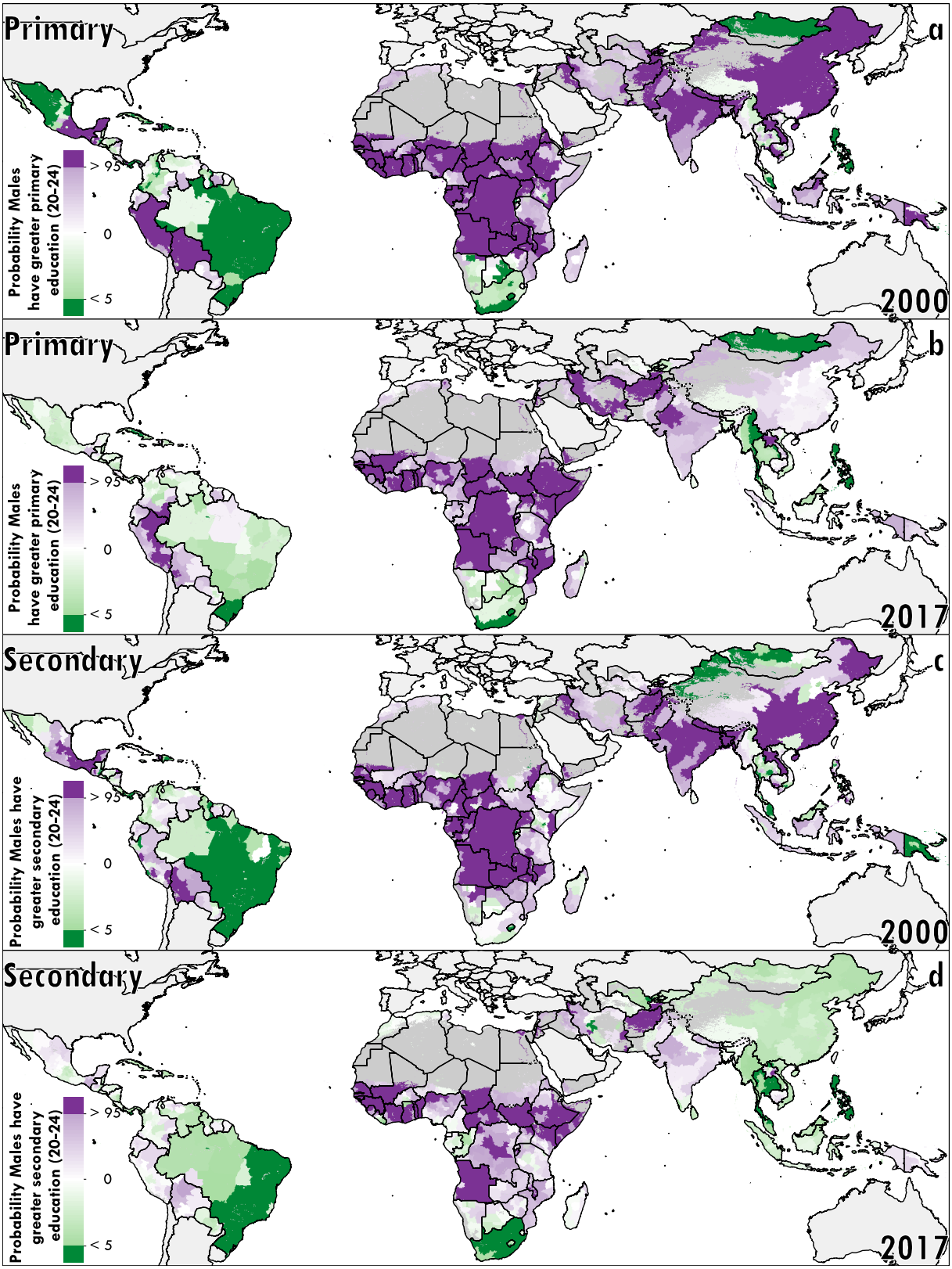
Correspondence and requests for materials should be addressed to S.I.H.

Peer review information *Nature* thanks M. Dolores Ugarte and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

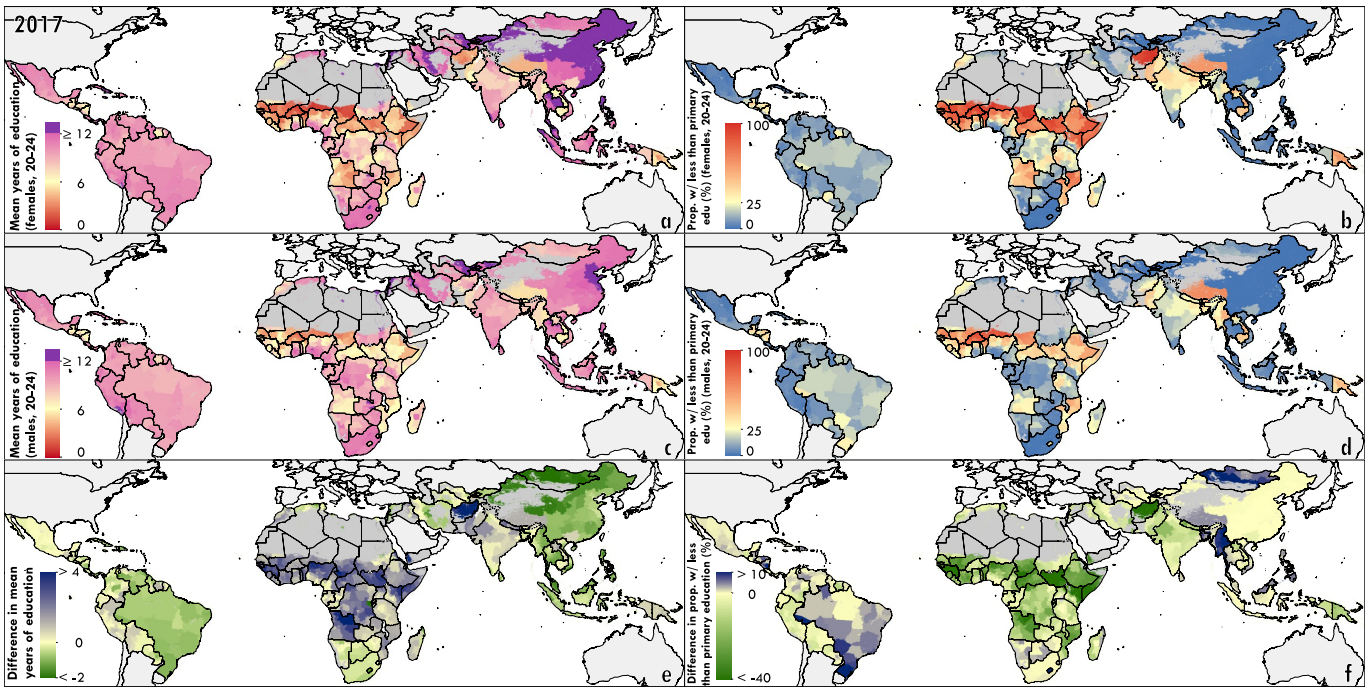
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Modelling regions based on geographical and SDI regions from the GBD. Modelling regions were defined as follows. Andean South America, Central America and the Caribbean, central sub-Saharan Africa, East Asia, eastern sub-Saharan Africa, Middle East, North Africa, Oceania, Southeast Asia, South Asia, southern sub-Saharan Africa, Central Asia, Tropical South America, and western sub-Saharan Africa. Regions in grey were not included in our models due to high-middle and high SDIs[27]. The map was produced using ArcGIS Desktop 10.6.



Extended Data Fig. 2 | Probability that the ratio of men to women aged 20–24 years who attained primary and secondary education is >1 in 2000 and 2017. a–**d**, Probability that ratio is >1 (for example, men complete at a higher rate than women) for attaining primary education (**a**, **b**) and secondary education (**c**, **d**), aggregated to first administrative-level units in 2000 (**a**, **c**) and 2017 (**b**, **d**). Maps were produced using ArcGIS Desktop 10.6.



Extended Data Fig. 3 | Average educational attainment and proportion with no primary school at the first administrative level and absolute difference between women and men aged 20–24 years. a–d, Average educational attainment for women (**a**) and men (**c**) and proportion with no primary school for women (**b**) and men (**d**) aged 20–24 years in 2017. **e**, **f**, The absolute difference in average educational attainment between men and women aged 20–24 years in 2017 (**e**) and proportion of individuals with no primary school education (**f**). Maps reflect administrative boundaries, land cover, lakes and population; grey-coloured grid cells were classified as 'barren or sparsely vegetated' and had fewer than ten people per 1 × 1-km^2 grid cell[49,58–60,62], or were not included in these analyses. Interactive visualization tools are available at https://vizhub.healthdata.org/lbd/education. Maps were produced using ArcGIS Desktop 10.6.

nature research

Corresponding author(s): Simon I.Hay and Nick Graetz

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☐	☒	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No primary data collection was carried out for this analysis.
Data analysis	This analysis was carried out using R version 3.5.0. The main geostatistical models were fit using R-INLA version 18.07.12. All code used for these analyses is publicly available online at http://ghdx.healthdata.org/.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The findings of this study are supported by data that are available in public online repositories, data that are publicly available upon request from the data provider, and data that are not publicly available due to restrictions by the data provider and which were used under license for the current study. A detailed table of data

sources and availability can be found in the Supplemental Information.

Administrative boundaries were retrieved from the Global Administrative Unit Layers (GAUL) dataset, implemented by FAO within the CountrySTAT and Agricultural Market Information System (AMIS) projects. Land cover was retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota. Lakes were retrieved from the Global Lakes and Wetlands Database (GLWD), courtesy of the World Wildlife Fund and the Center for Environmental Systems Research, University of Kassel. Populations were retrieved from WorldPop.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was calculated as the number of unique data source location-pairs with observations of years of educational attainment for males and females. This sample size is reported in the methods section as "We compiled a database of survey and census datasets that contained geocoding of subnational administrative boundaries or GPS coordinates for sampled clusters. These included datasets from 528 sources (see Supplementary Table 2).
Data exclusions	As described n the methods section of the main text (with greater detail in the Supplementary Information) we did not include data from Western Sahara or French Guiana, due to lack of availability. 42 of 105 included countries have only subnational administrative level data. Employing a Bayesian model-based geostatistical framework and synthesizing geolocated data from 517 household and census datasets, this analysis provides subnational estimates of mean years of education and proportion of the population attaining key levels of education for women of reproductive age (15-49), women age 20-24, and equivalent male age-bins between 2000-2017 in low- and middle-income countries. This includes 105 countries across all low- and middle-income countries. Countries were selected for inclusion in this analysis using the Socio-demographic Index (SDI) published in the GBD46. The SDI is a measure of development that combines education, fertility, and poverty. Countries in the Middle, Lower-Middle, or Low SDI quintiles were included, with several exceptions. Albania, Bosnia, and Moldova were excluded despite Middle SDI status due to geographic discontinuity with other included countries and lack of available survey data. Libya, Malaysia, Panama, and Turkmenistan were included despite Higher-Middle SDI status to create better geographic continuity. We do not estimate for American Samoa, Federated States of Micronesia, Fiji, Kiribati, Marshall Islands, Samoa, Solomon Islands, or Tonga, where no available survey data could be sourced.
Replication	This is an observational study using many years of survey and surveillance data and could be replicated.
Randomization	This analysis is an observational mapping study and there were no experimental groups.
Blinding	Blinding was not relevant to this study, as it was an observational study using survey and surveillance data.

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Internal state dynamics shape brainwide activity and foraging behaviour

https://doi.org/10.1038/s41586-019-1858-z

Received: 7 December 2018

Accepted: 18 November 2019

Published online: 18 December 2019

João C. Marques[1,2,4], Meng Li[1,3,4], Diane Schaak[1], Drew N. Robson[1,3*] & Jennifer M. Li[1,3*]

The brain has persistent internal states that can modulate every aspect of an animal's mental experience[1–4]. In complex tasks such as foraging, the internal state is dynamic[5–8]. *Caenorhabditis elegans* alternate between local search and global dispersal[5]. Rodents and primates exhibit trade-offs between exploitation and exploration[6,7]. However, fundamental questions remain about how persistent states are maintained in the brain, which upstream networks drive state transitions and how state-encoding neurons exert neuromodulatory effects on sensory perception and decision-making to govern appropriate behaviour. Here, using tracking microscopy to monitor whole-brain neuronal activity at cellular resolution in freely moving zebrafish larvae[9], we show that zebrafish spontaneously alternate between two persistent internal states during foraging for live prey (*Paramecia*). In the exploitation state, the animal inhibits locomotion and promotes hunting, generating small, localized trajectories. In the exploration state, the animal promotes locomotion and suppresses hunting, generating long-ranging trajectories that enhance spatial dispersion. We uncover a dorsal raphe subpopulation with persistent activity that robustly encodes the exploitation state. The exploitation-state-encoding neurons, together with a multimodal trigger network that is associated with state transitions, form a stochastically activated nonlinear dynamical system. The activity of this oscillatory network correlates with a global retuning of sensorimotor transformations during foraging that leads to marked changes in both the motivation to hunt for prey and the accuracy of motor sequences during hunting. This work reveals an important hidden variable that shapes the temporal structure of motivation and decision-making.

We placed freely moving zebrafish larvae in a behavioural arena (35 mm in diameter) with *Paramecia* (Fig. 1, Methods), and simultaneously recorded behaviour and whole-brain neural activity at cellular resolution for 50 to 80 min. We continuously tracked fish position and heading, eye orientation, tail curvature and *Paramecia* positions (Fig. 1a, Extended Data Fig. 1a–c, Supplementary Video 1, Methods). Each motor movement is classified into one of seven observed movement types[10,11] (Extended Data Fig. 1d–i, Supplementary Video 2, Methods). We manually annotated the start, end and outcome of each hunting attempt. Each hunting sequence begins with eye convergence (Extended Data Fig. 1b, c), in which the eyes—which are ordinarily parallel—reorient towards the midline to enable prey pursuit[12]. The animal then pursues the prey using high-precision turns (J-turns) and swims (approach swims)[13], and attempts to ingest the prey with capture swims[14], J-turns, high-angle turns or routine turns (Extended Data Fig. 2a–f). Every hunting sequence ends with one of four possible outcomes—success, miss, abort or reject (Extended Data Fig. 2a, d–f, Supplementary Video 1, Methods).

To investigate the overall dynamics of the behaviour of the animal across time, we defined a metric for exploratory drive as the diameter of the smallest circle enclosing the animal's trajectory in 5-s bins, which we call dispersal distance (Fig. 1a). We fit a hidden Markov model (HMM) to dispersal distance to define two behavioural states (Fig. 1b–d, Methods). Dispersal distance has a bimodal distribution (Extended Data Fig. 3a), with substantially higher dispersal in the exploration state (9.6 ± 2.5 mm, mean ± s.d., $n = 36$ animals) than in the exploitation state (2.3 ± 1.3 mm, $n = 36$ animals). On average, the exploitation state persists for 7.1 ± 3.9 min, and exploration state persists for 5.5 ± 6.2 min (mean ± s.d., $n = 36$ animals). The duration of each state follows an exponential distribution ($\lambda = 0.17\ \text{min}^{-1}$ for exploitation, $\lambda = 0.22\ \text{min}^{-1}$ for exploration, pooled distribution of $n = 36$ animals) (Fig. 1d, Extended Data Fig. 3g).

To independently confirm our choice of behavioural states, we assembled a comprehensive set of ten behavioural parameters, such as the rate of each movement type, and the rate of hunting initiations and outcomes (Methods). Though some individual behavioural parameters may be continuously distributed across the HMM states (Extended Data Fig. 3b, c), principal component analysis (PCA) of all behavioural parameters (Fig. 1e, Extended Data Fig. 3d–f, Methods) confirms the existence of

[1]Rowland Institute at Harvard University, Cambridge, MA, USA. [2]Present address: Champalimaud Research, Champalimaud Centre for the Unknown, Lisbon, Portugal. [3]Present address: Max Planck Institute for Biological Cybernetics, Tübingen, Germany. [4]These authors contributed equally: João C. Marques, Meng Li. *e-mail: drew.robson@tuebingen.mpg.de; jennifer.li@tuebingen.mpg.de

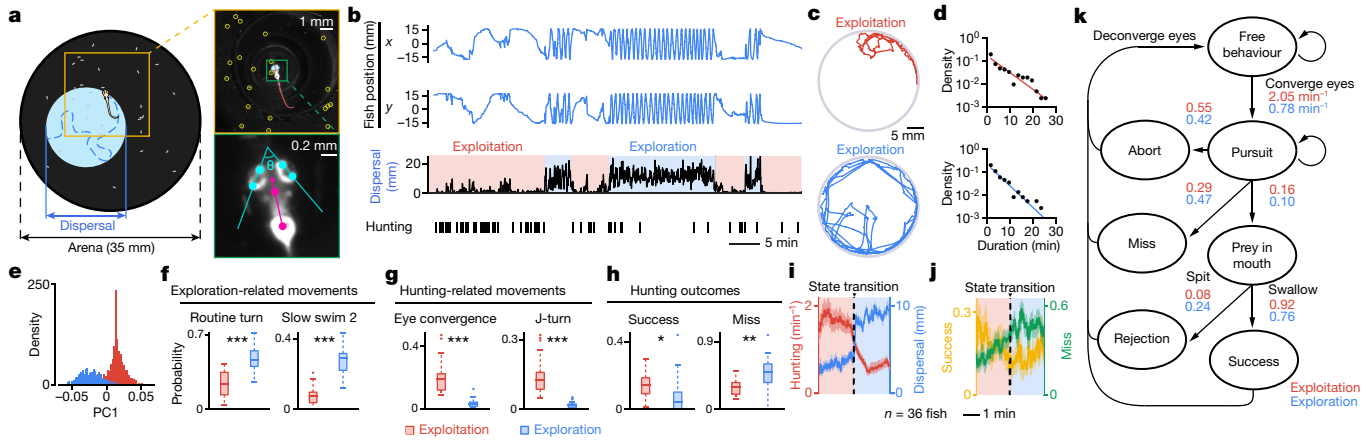


Fig. 1 | Freely moving zebrafish larvae display behavioural-state oscillations during foraging. a, Zebrafish larvae hunt for *Paramecia* in a behavioural arena (left) while we track the prey (top right, yellow), fish tail curvature (top right, red), heading vector (bottom right, magenta) and eye orientation (bottom right, cyan). Dispersal distance is defined as the diameter of the smallest circle encompassing the trajectory of the fish over 5 s. **b**, An example of fish position and dispersal distance across one hour, representative of $n = 36$ animals. Shaded regions depict exploitation state (red) and exploration state (blue) inferred by a HMM (Methods). A raster of hunting events is shown (bottom, black ticks). **c**, Representative 5-min movement trajectories in the behavioural arena during exploitation state (top) and exploration state (bottom). **d**, State duration distributions for exploitation (top) and exploration (bottom) ($n = 36$ animals), measured in two-minute bins (black dots) and a maximum likelihood exponential model fit (red and blue lines) to the underlying duration data. **e**, PCA was performed using ten behavioural variables (Methods). The histogram shows the distribution of the first principal component (PC1) colour-coded by whether the observation was classified as exploration (blue) or exploitation (red) by the HMM. **f**–**h**, State-dependent differences in the conditional probability (Extended Data Fig. 2t) of exploration-related movements (**f**), hunting-related movements (**g**) and hunting outcomes (**h**). $^*P < 0.05$, $^{**}P < 0.01$, $^{***}P < 0.001$. $n = 25$ animals, 2,972 hunting attempts and 88,814 movements. **i**, **j**, Time-aligned averages (mean ± s.e.m.) of behavioural metrics after aligning to HMM-inferred transitions from exploitation to exploration, including dispersal distance (**i**, blue), hunting rate (**i**, red), conditional probability of success per hunting initiation (**j**, orange) and conditional probability of miss per hunting initiation (**j**, teal) ($n = 36$ animals). **k**, Hunting behaviour as a Markov decision process. Nodes represent decision points in the hunting sequence. Arrows show state-dependent transition probability in exploitation state (red) and exploration state (blue).

two opposing behavioural states. Most of the behavioural variance was explained by the first principal component (Extended Data Fig. 3e), which has a bimodal distribution that matches the hidden Markov states inferred from dispersal distance (Fig. 1e, Extended Data Fig. 3d).

The exploration and exploitation states define an axis for global changes in behaviour (Fig. 1f–k, Extended Data Figs. 2, 3), including hunting initiation rate (Fig. 1g, i, k), the conditional probability of hunting success per hunting initiation (Fig. 1h, j), the Markov transition

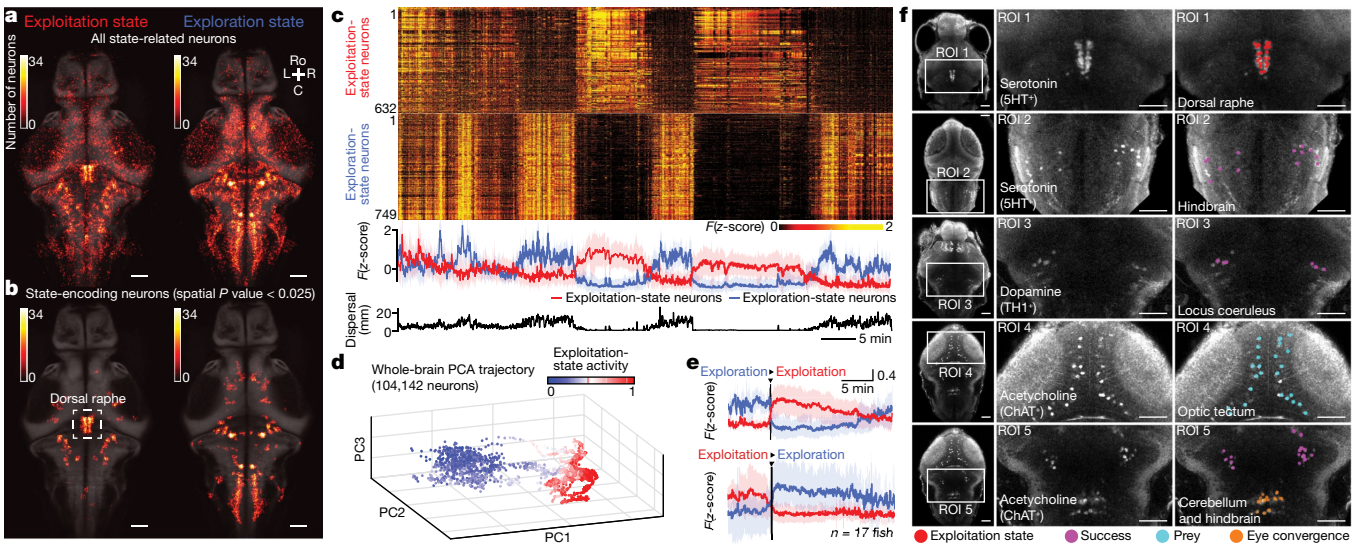


Fig. 2 | Opposing oscillatory networks encode the exploitation and exploration states. a, Locations of all neurons encoding exploitation state (left) or exploration state (right), projected across 17 coregistered animals. Ro, rostral; C, caudal; L, left; R, right. **b**, State-encoding neurons with significant spatial colocalization across animals (spatial P value < 0.025) (Methods). **c**, Top, activity raster of individual exploitation-state-encoding and exploration-state-encoding neurons from a representative animal. Middle, average activity (mean ± s.d.) of exploitation-state-encoding (red) and exploration-state-encoding (blue) neurons. Activity is shown as z-scored fluorescence (Methods). Bottom, dispersal distance across time. **d**, PCA trajectory of whole-brain activity (104,142 neurons) from a representative animal, colour-coded by the activity of exploitation-state-encoding neurons. **e**, Average activity (mean ± s.d.) of exploitation-state-encoding (red) and exploration-state-encoding (blue) neurons across the transition from exploration to exploitation (top), and the transition from exploitation to exploration (bottom) ($n = 17$ animals). **f**, Left, representative transverse sections from whole-brain immunostaining for serotonin (rows 1 and 2), dopamine (row 3) and acetylcholine (rows 4 and 5). Middle and right, expanded view of each region of interest (ROI) showing antibody-positive neurons encoding exploitation state (red), success (magenta), prey angle (cyan) and eye convergence (orange). Scale bars, 50 µm.

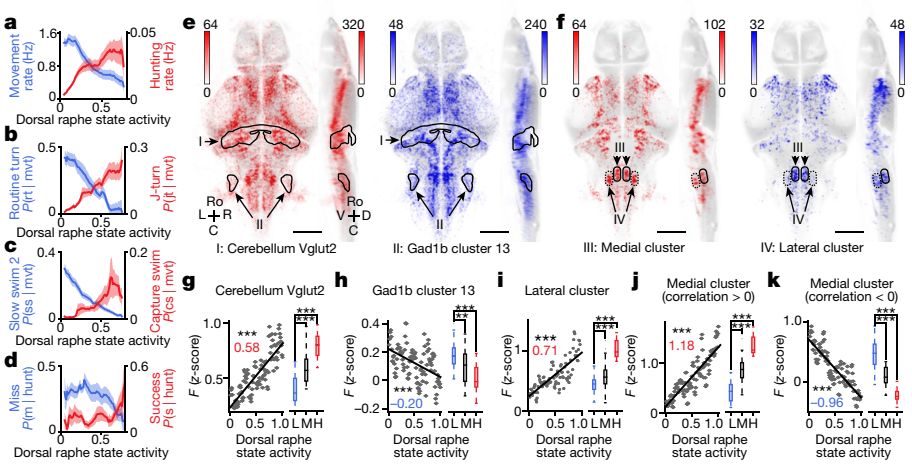


Fig. 3 | Neuromodulation of sensorimotor activity by dorsal raphe state neurons. a–**d**, Normalized dorsal raphe state activity is correlated with movement rate (**a**, blue), hunting rate (**a**, red), conditional probability per movement of routine turns (**b**, blue), J-turns (**b**, red), type 2 slow swims (**c**, blue) and capture swims (**c**, red), and conditional probability per hunting initiation of success (**d**, red) and miss (**d**, blue). Plots show mean ± s.e.m., $n = 16$ animals. **e**, **f**, Spatial density maps of sensorimotor neurons (**e**) and eye-convergence neurons (**f**) with positive (left, red) or negative (right, blue) correlation to normalized dorsal raphe state activity across all hunting events. Colour maps indicate total number of neurons after projection across 17 animals. Colour bars in **e**, **f** indicate number of neuronal centroids. V, ventral; D, dorsal. **g**–**k**, Left, event-triggered activity (average from 0 s to +2 s from hunting initiation) as a function of normalized dorsal raphe state activity. Slope of the regression fit is shown on the plot. $^{**}P < 0.01$, $^{***}P < 0.001$; two-sample t-test. Right, box plot of the same event-triggered activity binned into low (< 0.3, L), middle ($0.3 \leq x \leq 0.7$, M) and high (> 0.7, H) normalized dorsal raphe state activity. Box plots throughout depict the median (horizontal line), first and third quartiles (box) and the data range within 1.5-fold of the interquartile range of the first and third quartiles, respectively. Dorsal raphe state activity has a positive relationship with activity of neurons in the cerebellum (**g**), eye convergence neurons in two lateral hindbrain clusters (**i**) and a subpopulation of eye convergence neurons in two medial-hindbrain clusters (**j**). Dorsal raphe state activity has a negative relationship with the activity of neurons in the Gad1b cluster 13 region (**h**) and a subpopulation of eye convergence neurons in two medial hindbrain clusters (**k**). Dorsal raphe state activity was normalized from 0 to 1 for each animal to facilitate comparisons of neuromodulation across animals (Methods).

probabilities throughout the hunting sequence (Fig. 1k), and the conditional probability of each movement type (Fig. 1f, g, Extended Data Figs. 2t, 3c).

During exploration, an increase in overall movement rate and the probability of high-speed turns (routine turns) and swims (type 2 slow swims) (Fig. 1f, Extended Data Fig. 3c) combine to promote spatial dispersal. By contrast, during exploitation, the probability of hunting-related movements such as eye convergence, J-turns, approach swims and capture swims significantly increases (Fig. 1g, Extended Data Figs. 2t, 3c). The increased probability of hunting sequences that culminate in a capture swim in the exploitation state (Extended Data Fig. 2s) contributes to increased hunting success (Extended Data Fig. 2g, j, m, p). There is a strong inverse relationship between movement rate and the probability of hunting-related movements (Extended Data Figs. 2u, 3c).

We found that behavioural-state transitions are not solely dependent on changes in satiety. Larvae placed in a behavioural arena without prey, which prevents fluctuations in satiety owing to prey consumption, continue to exhibit clear alternations in global dispersal and local search (Extended Data Fig. 4a–d). This suggests an intrinsic mechanism that can drive spontaneous oscillations in motivation and decision-making during foraging. Next, we compared larvae that were fed until the start of the experiment with siblings that were starved for 24 h (Extended Data Fig. 4e, f). Starved animals spent more time in the exploitation state for the first 10 min of the experiment, but this effect was not statistically significant (Extended Data Fig. 4e). Thus, although it is possible that hunger may weakly modulate exploitation- and exploration-state transitions, more comprehensive studies are required to fully address this relationship.

Brainwide encoding of internal state

We imaged whole-brain calcium activity at cellular resolution in freely swimming zebrafish larvae expressing pan-neuronal H2B-GCaMP6s[15] during foraging, and extracted the fluorescence across time $F(t)$ for each neuronal centroid in the brain using non-negative matrix factorization (NMF)[16] (Methods). Using linear regression (Methods), we identified neural populations showing correlation to prey detection (Extended Data Fig. 5f, i), eye convergence (Extended Data Fig. 5g), prey ingestion (Extended Data Fig. 5h), each movement type (Extended Data Fig. 5j) and internal state (421 ± 141 exploitation-state neurons and 910 ± 319 exploration-state neurons per animal, mean ± s.d.) (Fig. 2a–e).

For each neural population, we applied a spatial P value test to identify functional clusters that are spatially consistent across animals (Extended Data Fig. 5a, Methods). In particular, the dorsal raphe had the highest spatial enrichment of exploitation-state-related neurons among all brain regions (Fig. 2a, b, Extended Data Fig. 5b–e, Supplementary Video 3). Approximately 19 ± 8% (mean ± s.d.) of dorsal raphe neurons encode the exploitation state.

By PCA analysis of whole-brain activity[17], exploitation-state neurons define the major axis of global brain activity (Fig. 2d, Extended Data Fig. 6). The peak activity of exploitation-state neurons is predictive of exploitation-state duration (Extended Data Fig. 7a). The activity of these neurons is notably persistent and asymmetric, peaking rapidly and decaying slowly (Fig. 2e). Transition from exploitation to exploration occurs when the activity of these neurons relaxes back to baseline (Fig. 2c, e).

Because some neurons with apparent correlations to exploitation state might be attributable to transient, event-locked activity related to prey pursuit or ingestion, we excluded increasingly large time intervals encompassing all hunting events across the entire experiment (Methods). We then repeated the linear regression analysis to confirm that exploitation-state neuron activity is persistently and stably correlated with behavioural state and not merely due to transient activity around individual hunting events (Extended Data Fig. 8a, b).

We took two further steps to validate the dorsal raphe exploitation-state neurons. First, we were able to independently discover these

neurons using unsupervised k-means clustering ($k = 5$) on the pairwise activity correlation matrix between all dorsal raphe neurons (Extended Data Fig. 8c, Methods). In every fish ($n = 17$), we found that the cluster with the highest average intra-cluster correlation encodes the exploitation state (Extended Data Fig. 8d). Second, we found that the mean correlation between behavioural state and the activity of dorsal raphe state neurons in the same animal was 0.468, whereas the correlation between the behavioural state of one animal and the exploitation-state activity of another was 0.024 (Extended Data Fig. 8f). This difference is highly significant ($P < 10^{-8}$, two-sample t-test), demonstrating that the discovery of dorsal raphe state neurons is not a result of spurious correlations[18,19].

Exploration-associated neurons differ in several important ways. Whereas exploration-associated neurons show consistently elevated activity during exploration, they do not show asymmetric rise and decay kinetics (Fig. 2c, e). Notably, the peak activity of these neurons is not predictive of the duration of the exploration state (Extended Data Fig. 7a), which suggests that they may be primarily involved in the expression of exploration behaviour, but not in determining the duration of the exploration state.

To determine the neuromodulators that participate in exploitation and exploration, we examined the activity of serotonergic, dopaminergic and acetylcholinergic neurons throughout the entire brain using the MultiMAP method[20,21] (Fig. 2f, Extended Data Fig. 9, Supplementary Video 4, Methods). Each neuromodulatory population has distinct roles in internal state and hunting, is heterogeneously composed of multiple functional types and exhibits distinct temporal dynamics (Fig. 2f, Extended Data Fig. 10).

Serotonergic neurons in the dorsal raphe have persistent activity that can last for minutes and correlates with exploitation state (Fig. 2f, Extended Data Fig. 10a, d). Serotonergic neurons in the dorsal hindbrain (Fig. 2f, Extended Data Fig. 10a, d), dopaminergic neurons in the locus coeruleus and medulla oblongata[22] (Fig. 2f, Extended Data Fig. 10b, e), as well as acetylcholinergic neurons in the cerebellum, are preferentially activated by hunting success (Fig. 2f, Extended Data Fig. 10c, f). Tectal acetylcholinergic neurons are activated when prey enter their receptive fields (Fig. 2f, Extended Data Fig. 10f), and the medial rhombencephalic acetylcholinergic population correlates strongly with eye convergence (Fig. 2f, Extended Data Fig. 10c).

State-dependent sensorimotor tuning

We find that specific sensorimotor clusters are modulated by internal state (Fig. 3, Methods). Dorsal raphe state activity correlates with significant changes in locomotion, hunting initiation and accuracy (Fig. 3a–d). Consistent with this observation, dorsal raphe state activity is positively correlated with the activity of the cerebellum Vglut2 cluster[23] (Fig. 3e, g)—which encodes eye convergence and prey ingestion (Extended Data Fig. 5g, h)—and negatively correlated with the activity of Gad1b cluster 13 in the hindbrain[23] (Fig. 3e, h), which encodes routine turns (Extended Data Fig. 5j).

Moreover, even within a functional network that encodes the same behavioural event (such as eye convergence or prey ingestion) (Extended Data Fig. 5g, h), the degree of correlation with dorsal raphe activity varies across different anatomical modules (Fig. 3f, i–k). In rhombomere 6, two lateral clusters contain more than 14.1-fold more neurons that are positively modulated by state (Fig. 3f, i), whereas two medial clusters contain a sizable number of neurons that are both positively and negatively modulated by state (366 and 145 neurons, respectively) (Fig. 3f, j–k).

Our results suggest that state-dependent neuromodulation targets specific sensorimotor modules in the brain. The diversity of neuromodulatory targets and effects suggests a global retuning of sensorimotor transformations as a function of internal state.

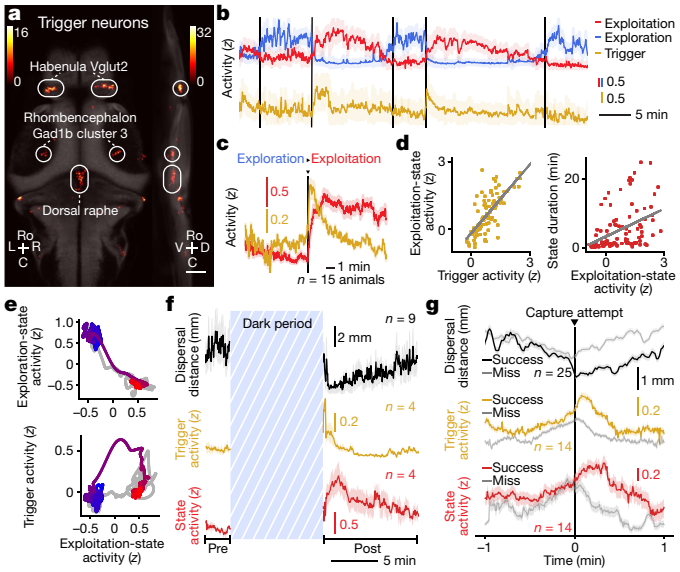


Fig. 4 | A network of trigger neurons is active at state transition from exploration to exploitation. a, Spatial map of trigger neurons that are active at the transition from exploration to exploitation state ($n = 15$ animals, all neurons shown have a spatial P value < 0.025). Colour bar indicates number of neuronal centroids. **b**, Example fluorescent activity (mean ± s.d.) of dorsal raphe exploitation-state neurons (red), exploration-state neurons (blue), and trigger neurons (yellow) across multiple state oscillations (vertical black lines). **c**, Mean activity of dorsal raphe exploitation-state neurons (red) and trigger neurons (yellow) aligned to state transition from exploration to exploitation. **d**, Peak activity of trigger neurons scales linearly with the peak activity of dorsal raphe exploitation-state neurons (left). Peak activity of dorsal raphe exploitation-state neurons correlates with the duration of exploitation state (right). **e**, Phase space plot of exploitation- and exploration-state activity (top). Phase space plot of dorsal raphe state activity and trigger activity (bottom). Both plots depict one cycle based on average state-aligned activity. Blue-to-red colour map depicts the transition from exploration to exploitation state. The return trajectory from exploitation to exploration state is shown in grey. **f**, Zebrafish larvae were placed in the presence of *Paramecia* and imaged for 50 min in light, 10 min in darkness (all light sources including fluorescence excitation source were turned off), followed by 20 min in light. The 10 min of darkness followed by light strongly induces the exploitation state, as shown by the persistent suppression of dispersal distance (black) and activation of raphe state neurons (red) and trigger neurons (yellow). **g**, Successful ingestion of prey weakly promotes the exploitation state, as shown by suppression of dispersal distance, weak activation of trigger neurons and weak activation of dorsal raphe state neurons after successful capture attempts (black, yellow, and red, respectively) compared with missed capture attempts (grey). Data in **b**, **c**, **f**, **g** are mean ± s.e.m.

Neural dynamics at state transitions

We find, by linear regression, a distributed network of neurons that are triggered specifically at the state transition from exploration to exploitation (Fig. 4, Methods). Spatially significant clusters are primarily located in the ventrolateral habenula, dorsal raphe and Gad1b clusters in the rhombencephalon (Fig. 4a–c, Extended Data Fig. 5b–e, Supplementary Video 3). The enrichment of the trigger neurons in the habenula, which makes both direct and indirect synaptic connections to the dorsal raphe[24–26], suggests a possible circuit mechanism for internal state transitions.

Neurons in this trigger network are weakly activated before state transition and peak at the time of the state transition (Fig. 4c). The peak activity amplitude of this trigger network is linearly correlated with the peak activation of dorsal raphe state neurons, which is in turn correlated with the duration of the exploitation state (Fig. 4d).

The short impulsive dynamics of the trigger neurons and the persistent asymmetric dynamics of dorsal raphe state neurons (Fig. 4c) can be combined to form a two-component nonlinear dynamical system that produces a nonsinusoidal repetitive output (Fig. 4e, Extended Data Fig. 7). A stochastic dynamical system model can recapitulate the distribution of behavioural state durations and the dynamics of trigger and state neuron activity (Extended Data Fig. 7e, f, Methods).

The trigger network can be either spontaneously active or driven by sensory stimuli (Fig. 4f, g, Extended Data Fig. 7b–d). For instance, 10 minutes of darkness followed by light strongly suppresses dispersal distance for minutes, triggers transition into the exploitation state in all animals (9 out of 9), activates the trigger network and induces a prolonged activation of dorsal raphe state neurons (Fig. 4f). In addition, following successful prey capture, we also observe a momentary decrease in spatial dispersal, a weak activation of the trigger network and a brief increase in dorsal raphe state activity (Fig. 4g). This could potentially act as a mechanism to prolong the exploitation state in an advantageous environment.

Our results demonstrate that zebrafish larvae continuously oscillate between opposing global brain states during foraging, and that a dorsal raphe subpopulation may coordinate a fundamental trade-off between overall motor activity and motor bias towards hunting. This subpopulation encodes a generalized motivational state—not simply hunger or satiety[27]—and may provide an evolutionary link in the neural mechanism of internal state oscillations from *C. elegans*[5] to rodents[6] and primates[2,28].

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1858-z.

1. Dayan, P. How to set the switches on this thing. *Curr. Opin. Neurobiol.* **22**, 1068–1074 (2012).
2. Aston-Jones, G. & Cohen, J. D. An integrative theory of locus coeruleus-norepinephrine function: adaptive gain and optimal performance. *Annu. Rev. Neurosci.* **28**, 403–450 (2005).
3. Fox, M. D. et al. The human brain is intrinsically organized into dynamic, anticorrelated functional networks. *Proc. Natl Acad. Sci. USA* **102**, 9673–9678 (2005).
4. Nair, J. et al. Basal forebrain contributes to default mode network regulation. *Proc. Natl Acad. Sci. USA* **115**, 1352–1357 (2018).
5. Flavell, S. W. et al. Serotonin and the neuropeptide PDF initiate and extend opposing behavioral states in *C. elegans*. *Cell* **154**, 1023–1035 (2013).
6. Lottem, E. et al. Activation of serotonin neurons promotes active persistence in a probabilistic foraging task. *Nat. Commun.* **9**, 1000 (2018).
7. Cohen, J. D., McClure, S. M. & Yu, A. J. Should I stay or should I go? How the human brain manages the trade-off between exploitation and exploration. *Philos. Trans. R. Soc. B* **362**, 933–942 (2007).
8. Charnov, E. L. Optimal foraging, the marginal value theorem. *Theor. Popul. Biol.* **9**, 129–136 (1976).
9. Kim, D. H. et al. Pan-neuronal calcium imaging with cellular resolution in freely swimming zebrafish. *Nat. Methods* **14**, 1107–1114 (2017).
10. Marques, J. C., Lackner, S., Félix, R. & Orger, M. B. Structure of the zebrafish locomotor repertoire revealed with unsupervised behavioral clustering. *Curr. Biol.* **28**, 181–195.e5 (2018).
11. Marques, J. C. & Orger, M. B. Clusterdv: a simple density-based clustering method that is robust, general and automatic. *Bioinformatics* **35**, 2125–2132 (2019).
12. Bianco, I. H. & Engert, F. Visuomotor transformations underlying hunting behavior in zebrafish. *Curr. Biol.* **25**, 831–846 (2015).
13. McElligott, M. B. & O'malley, D. M. Prey tracking by larval zebrafish: axial kinematics and visual control. *Brain Behav. Evol.* **66**, 177–196 (2005).
14. Borla, M. A., Palecek, B., Budick, S. & O'Malley, D. M. Prey capture by larval zebrafish: evidence for fine axial motor control. *Brain Behav. Evol.* **60**, 207–229 (2002).
15. Freeman, J. et al. Mapping brain activity at scale with cluster computing. *Nat. Methods* **11**, 941–950 (2014).
16. Friedrich, J. et al. in *NIPS Workshop on Statistical Methods for Understanding Neural Systems* https://pdfs.semanticscholar.org/e4ff/845a4b996482f4ef491fff4581a59d949800.pdf (2015).
17. Niessing, J. & Friedrich, R. W. Olfactory pattern classification by discrete neuronal network states. *Nature* **465**, 47–52 (2010).
18. Granger, C. W. J. in *A Companion to Theoretical Econometrics* (ed. Balgati, B. H.) Ch. 26 (Blackwell, 2007).
19. Phillips, P. C. B. Understanding spurious regressions in econometrics. *J. Econom.* **33**, 311–340 (1986).
20. Lovett-Barron, M. et al. Ancestral circuits for the coordinated modulation of brain state. *Cell* **171**, 1411–1423 (2017).
21. Marquart, G. D. et al. High-precision registration between zebrafish brain atlases using symmetric diffeomorphic normalization. *Gigascience* **6**, 1–15 (2017).
22. Kastenhuber, E., Kratochwil, C. F., Ryu, S., Schweitzer, J. & Driever, W. Genetic dissection of dopaminergic and noradrenergic contributions to catecholaminergic tracts in early larval zebrafish. *J. Comp. Neurol.* **518**, 439–458 (2010).
23. Randlett, O. et al. Whole-brain activity mapping onto a zebrafish brain atlas. *Nat. Methods* **12**, 1039–1046 (2015).
24. Bianco, I. H. & Wilson, S. W. The habenular nuclei: a conserved asymmetric relay station in the vertebrate brain. *Philos. Trans. R. Soc. B* **364**, 1005–1020 (2009).
25. Amo, R. et al. Identification of the zebrafish ventral habenula as a homolog of the mammalian lateral habenula. *J. Neurosci.* **30**, 1566–1574 (2010).
26. Kalén, P., Karlson, M. & Wiklund, L. Possible excitatory amino acid afferents to nucleus raphe dorsalis of the rat investigated with retrograde wheat germ agglutinin and d-[^{3}H] aspartate tracing. *Brain Res.* **360**, 285–297 (1985).
27. Filosa, A., Barker, A. J., Dal Maschio, M. & Baier, H. Feeding state modulates behavioral choice and processing of prey stimuli in the zebrafish tectum. *Neuron* **90**, 596–608 (2016).
28. Fox, M. D. & Raichle, M. E. Spontaneous fluctuations in brain activity observed with functional magnetic resonance imaging. *Nat. Rev. Neurosci.* **8**, 700–711 (2007).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

Simultaneous behavioural and neural imaging by tracking microscopy

The behaviour arena is 35 mm in diameter and 810 μm in height, constructed according to the same design as a previous study[9]. In brief, the top and bottom of the chamber are made of glass to allow optical access from above and below. The chamber walls are cut from precast gas-permeable polydimethylsiloxane (PDMS; HT-6240, Rogers) using a computer-controlled blade (Curio, Silhouette).

We imaged whole-brain calcium activity at cellular resolution in freely swimming larval zebrafish expressing pan-neuronal H2B-GCaMP6s[15], a nuclear-localized reporter of neural activity. Tracking microscopy was performed as described in a previous study[9]. In brief, to stabilize the brain position of the fish during free behaviour, we use a motion cancellation system that consists of a NIR imaging setup for animal tracking, a behaviour arena, and *X*–*Y* motorized stages. The NIR imaging setup consists of a CMOS camera (BFS-U3-32S4M-C, FLIR), camera lens (13VM1040ASIR, Tamron), and optical filters (FES0900 and FEL0700, Thorlabs). The NIR camera collects images of the entire animal at 250 Hz with a FOV of 10 × 10 mm and a resolution of 608 × 608 pixels (16 μm per pixel). The behaviour arena is illuminated for dark field imaging by a custom ring light consisting of 60 narrow-angle 850 nm infrared LEDs (SFH4655-Z, Osram). Each image is precisely timed by driving the LEDs with a short pulse (10–100 μs) that is synchronized with the camera exposure. The pulse driver circuit consists of a MOSFET (PSMN4R3-30PL,127, NXP Semiconductors) and 10 current-limiting series resistors governing 10 strings of 6 LEDs. A pulse duration of 10–100 μs is used to avoid motion blur owing to rapid fish motion. The X-Y motorized stages (ILS200LM, Newport) consist of a stacked pair of three-phase ironless linear motors and a motor controller (XPS-Q8, Newport). A custom-machined open frame with high stiffness and low mass mechanically couples the behaviour arena to the stages and creates space underneath the arena for the NIR optics.

To record neural activity, we used differential illumination focal filtering (DIFF)[9] microscopy to image 1001 × 761 × 150 μm of the brain at cellular resolution at a volume rate of 2 Hz. Each volume consisted of 50 DIFF-sectioned images, computed from pairs of structured images collected at 200 Hz by an sCMOS camera (C13752-50U, Hamamatsu). The design and performance of the DIFF imaging system have been characterized in depth in a previous study[9].

Hunting assay

Six-to-eight-days-post-fertilization (dpf) larval zebrafish were held without food for at least 1 h before the start of the hunting assay. Immediately before the experiment, approximately 200–500 *Paramecia* were loaded into the behavioural arena. Then, a single 6–8-dpf zebrafish was placed in the behavioural arena for 1–2 h during the assay. The arena was illuminated from above by a custom circular array of 60 white LEDs (C512A-WNS-CZ0C0151, Cree)[9]. Hunting success rate shows no significant differences when assayed in stationary behavioural arenas or in the tracking microscope, as demonstrated in a previous study[9]. For the analysis in Fig. 4f and Extended Data Fig. 7b–d, a period of complete darkness was created by turning off the white LED array and the DIFF fluorescence microscope excitation source for 10 min.

High-resolution offline registration of fluorescent brain volumes for each animal

For each animal, we registered each fluorescent image from the tracking microscope to a high-resolution reference brain volume at micrometre resolution collected from the same animal. An in-depth description and characterization of the registration pipeline was published in a previous study[9]. In brief, an initial coarse registration is obtained by optimizing a 3D rigid transformation that maps the moving image to a plane within the reference brain volume. This planar surface is then finely subdivided into a deformable surface that is locally adjusted within the reference volume using a regularized piecewise affine transform.

Non-negative matrix factorization

Post-registration, we extracted fluorescence $F(t)$ across time t for each neuron in the brain using NMF[16] ($n = 17$ animals, 112,699 ± 10,370 neuronal centroids, mean ± s.d., per animal). NMF separates each neuron into two components—a spatial footprint and a time-varying activity component. We applied constrained NMF[16] to our whole-brain H2B-GCaMP6s datasets. NMF was performed independently for each axial section of a given brain volume. There can be double counting of centroids if a neuron spans more than one axial plane. Our axial sections are separated by 3 μm, and the average diameter of a zebrafish neuron is approximately 5 μm.

Registration to a common reference brain across animals

We used Advanced Normalization Tools (ANTs, v.2.2.0, running on an Ubuntu 16.04 Linux server)[21,29–31] and ITK-SNAP (v.3.6.0, running on a Windows desktop) to spatially map live fluorescent brain volumes from different animals into a common reference space (Z-Brain[23]). The Z-Brain atlas (Elavl3:H2B-RFP line) served as the fixed image, and live fluorescent brain volumes were used as the moving images. We first aligned the moving images to the fixed image by using ITK-SNAP to match the centre, orientation and scale of the brain. The aligned moving and fixed images were then saved in the NII file format. By combining linear (translation, rigid, similarity and affine, with 3, 6, 7 and 12 degrees of freedom, respectively) and nonlinear (diffeomorphic symmetric normalization algorithm, SyN) ANTs registration, we optimized the compromise between preservation of cell morphology and global alignment and achieved reasonably precise registration (within 1–2 cell diameters) of all live fluorescent brain volumes to the common reference. After testing a range of values for each of the ANTs registration parameters, we finally chose the following parameters: antsRegistration -verbose 1–dimensionality 3–float 1–output [moving_image, moving _image _Warped.nii.gz, moving _image_InverseWarped.nii.gz]–interpolation WelchWindowedSinc–use-histogram-matching 0–winsorize-image-intensities [0.005,0.995]–initial-moving-transform [fixed_image.nii, moving_image.nii, 1]–transform Rigid[0.1]–metric MI [fixed_image.nii, moving_image.nii, 1,32,Regular,0.25]–convergence [1000x500x250x0,1e-8,10]–shrink-factors 12x8x4x2–smoothing-sigmas 4x3x2x1vox–transform Affine[0.1]–metric MI [fixed_image.nii, moving_image.nii, 1,32,Regular,0.25]–convergence [1000x500x250x0,1e-8,10]–shrink-factors 12x8x4x2–smoothing-sigmas 4x3x2x1vox–transform SyN[0.1,6,0]–metric CC [fixed_image.nii, moving_image.nii, 1,2]–convergence [100x100x70x50x0,1e-7,10]–shrink-factors 12x8x4x2x1–smoothing-sigmas 5x3x2x1x0vox.

Registration of live fluorescent brain volume to immunostained brain volume in the same animal using ANTs

We applied a four-step procedure to register each live fluorescent brain volume to the same brain after fixation and immunostaining[21] (Extended Data Fig. 9). In all four steps, we used the immunostained brain as the fixed image and the live fluorescent brain as the moving image. The moving images were roughly aligned to the fixed image using ITK-SNAP (centre, orientation and scale), and then ANTs registration was applied.

As shown in Extended Data Fig. 9a, step 1 (global registration) was designed to obtain a global alignment between the live fluorescent brain volume and the immunostained brain volume. This step used the same ANTs parameters as in the previous section. The overall outline

of the registered live fluorescent brain volume was well-matched to the immunostained brain volume after step 1. However, because of the properties of the SyN nonlinear method, the global registration generates distortions in some regions of the live fluorescent brain volume, especially in the centre of the brain (for example, raphe). To optimize the quality of registration on both the global and local scale, we performed local registration using ANTs (steps 2–4). In step 2 (tiling), the fixed and moving images were subdivided into 12 tiles, such that each tile had a 50% overlap with adjacent tiles. In step 3 (local registration), we applied local registration to each of the 12 tiles, using the same ANTs parameters as the global registration. In step 4 (merging), all locally registered tiles were then merged together by alpha blending the overlapping regions.

In Extended Data Fig. 9b, we show the superimposed live fluorescent and immunostained brain volume before registration, after global registration and after global and local registration. In Extended Data Fig. 9c, we show the immunostained brain volume (green), live fluorescent brain volume (red), and the superimposed volume for the optic tectum (white box in Extended Data Fig. 9b) to further illustrate the registration performance (for example, note the precise overlap of the sparsely distributed neurons in the optic tectum, denoted by the white arrows). Post-registration, the correlation between live and immunostained brain volumes was 0.7 (Extended Data Fig. 9d), consistent with previous studies on high-precision alignment of brain volumes using ANTs[21].

Tail tracking

We used an optimization-based algorithm to trace the curvature of the fish tail. Specifically, for each NIR image of the fish, we created a high-resolution spline for the tail to enable further classification of movements into distinct movement types (see 'Classification of movements into bout types'). A sequential grid search was used to fit an optimized tail spline onto the fish tail. This spline consisted of 37 tail segments, each 80-µm long, parameterized by the local angle change at each segment boundary. This high-resolution optimization-based algorithm was specifically developed to achieve robust tail tracking in the presence of bright, highly dynamic prey in the vicinity of the tail.

Classification of movements into bout types

Movement bout detection. Larval zebrafish move in discrete bouts. Bout detection was performed by using a density valley clustering method to classify each motor movement into one of seven observed movement types (Extended Data Fig. 1d–h, Supplementary Video 2), as previously described[10].

The presence of noise in the optimization-based tail tracking can lead to false positives in the bout detection pipeline that do not correspond to tail movements. To make the bout detection pipeline more robust, we defined a set of diagnostic criteria to detect and exclude false positive bouts. Bouts with a minimum tail beat frequency smaller than 10 Hz came from failures to separately detect two adjacent bouts and were excluded. Bouts with maximum tail beat frequency larger than 100 Hz resulted from unstable tail tracking and were excluded as well. Bouts with maximum tail tip angle smaller than 5° resulted from tracking noise in the tip of the tail and were not included for further analysis. Bouts that had high values of maximum angular speed/bout maximum angle (> 33°ms^{-1}) or bout maximum angular speed (> 43°ms^{-1}) were confirmed to arise from upstream tail tracking errors and were excluded. Finally, all bouts were placed in the 73 kinematic parameter PCA space. The bouts that had a third principal component value (PC3) larger than 3 were outliers in this space. These bouts were confirmed to arise from tail tracking noise in stationary fish and were excluded.

Half-beat detection. Half-beats were detected as previously described[10].

Clustering and bout type classification. For each swim bout, we computed 73 kinematic parameters based on the bout and half-beat detection, as described previously in depth[10]. PCA was performed on the bout kinematic parameters on an initial set of 53 fish that were collected under hunting conditions with live prey (hunting), or under no-prey conditions (spontaneous swimming) in white light illumination. We embedded all the data into the PCA space and included for further analysis the first six principal components (PCs). We then separated the bout data by experimental condition (hunting or spontaneous swimming), the condition in which the fish were raised (raised with *Paramecia* or without), the behavioural state (exploration or exploitation), eye convergence (converged or unconverged), and in the case of the final hunting bout of each hunting sequence, the outcome of hunting (miss, abort, rejection or success). The datasets with a large number of points were still further subdivided into sets of 3,000 points. To find the number of clusters in each dataset, we used clusterdv[11]. To exclude outliers, the number of clusters was selected by drawing a rectangle on the decision plot (separability index versus local density).

The clusters were used to classify bout types as previously described in depth[10]. Bout types were classified for 28 foraging animals (1–2 h of behavioural recording per animal in the presence of *Paramecia*), and 7 nonforaging animals (about 2 h of behavioural recording per animal in the absence of *Paramecia*). Among these seven bout types, there are two types directly related to hunting events. Approach swims are small forward bouts that often occur when the prey is in front of the fish[10]. J-turns are high precision turns, which are composed of a series of unilateral J-bends (> 90°) at the caudal end of the tail (~20% from the tip of the tail)[12].

Validation of bout type classification. To validate the bout classification, we calculated—for each bout—12 kinematic parameters based on the fish position, heading and tracking of the tail. Displacement is the straight-line change in the head position per bout. Forward displacement corresponds to the forward component of the displacement vector. Maximum head yaw is the maximum peak-to-peak change in heading after the first bend and counter bend. The maximum bout angle is the magnitude of the largest heading change during a bout. Maximum angular speed is the maximum rate of change of heading during a bout. 'Tail maximum amplitude 2' is the maximum absolute tail bend amplitude, during a bout, at 0.6 mm from the swim bladder. 'Tail maximum amplitude 7' is the maximum absolute tail bend amplitude, during a bout, at 2.1 mm from the swim bladder. The C1 and C2 durations are the durations of the first and second half-beats. The C1 and C2 maximum angular speeds are the maximum angular velocities during the first and second half-beats. 'First and second tail maximum amplitudes 7' are the maximum absolute tail bend amplitudes, during the first and second half-beats, at 2.1 mm from the swim bladder.

Paramecia tracking

Paramecia appear as bright round regions of high intensity in the behaviour camera images. Therefore, the simplest strategy for tracking *Paramecia* is to estimate a static background image, subtract this background image from each image frame (separating *Paramecia* from background), perform intensity-based blob detection, segment the image into appropriately sized regions of interest representing *Paramecia*, and then extract the coordinates of each paramecium. Our implementation followed essentially these steps, but with a modification to the front end. During tracking microscopy, the background image for the behaviour camera is dynamic, primarily owing to axial motion of the imaging objective. Therefore, we replaced the static background subtraction step with a deep neural network[32–34] (DNN). In brief, we used DeeperCut[32] with default parameters to train a model based on a pretrained ResNet-50 network to take a raw camera image as input (that is, an image with *Paramecia* and dynamic background) and produce

an output image with high pixel intensity at only the *Paramecia* locations (an image with *Paramecia* and no background). Thus, the DNN performs the role of robust background subtraction, and produces a cleaner image that can be analysed by conventional intensity-based tracking (Fig. 1a, Extended Data Fig. 1a, Supplementary Video 1; see Supplementary Methods for details).

Eye position and orientation tracking

We also used the DNN approach for tracking the position, heading and eye orientation of the fish (Fig. 1a, Extended Data Fig. 1a, Supplementary Video 1). Six fish body parts were tracked by the DNN, consisting of front left eye, rear left eye, front right eye, rear right eye, centre of the brain and yolk. To generate the training dataset, 3,767 frames were randomly extracted from 11 recording sessions. The architecture of the DNN reduces the image resolution, such that a pixel in the score map corresponds to 8 × 8 pixels in the input frame. Thus, to obtain score maps at the same image resolution as the input frames, we first selected a 100 × 100 ROI in each frame, centred on the centre of the brain to cover all six of the tracked fish body parts. The ROIs were then up-sampled eightfold (800 × 800) to ensure the score maps had the same image resolution as the raw ROIs. The coordinates of the six fish body parts were manually labelled in each up-sampled ROI of the training data. The up-sampled ROIs of all training frames and the corresponding coordinates of the six fish body parts were then used to train the DNN. We used DeeperCut[32] with default parameters to train a model based on ResNet-50 for 1,030,000 iterations on a NVIDIA GeForce GTX 1080 Ti GPU for ~24 h.

We applied the well-trained DNN to all the recording sessions to calculate six score maps for each frame, corresponding to the six body parts. The resulting images were then processed using intensity-based tracking of each body part using a custom Julia implementation. To minimize the effects of tracking error, we summed all six score maps together to generate a global pattern of all the body parts for each frame. We then calculated the orientation of the fish and the eye convergence index for each frame. The orientations of the fish were defined as the orientations of the rays from the yolk to the centre of the brain of the fish. To calculate the eye convergence index, we first calculate the orientations of the eyes, defined as the angle of the ray from the rear eye to front eye relative to the heading vector of the fish. The eye convergence index was then defined as the difference of the individual eye angles. For all fish we recorded, there are two eye convergence states—converged and unconverged[12]. We use a threshold to define a putative start and end of each eye convergence event. To exclude false positives and accurately identify bouts in which the eyes were converged, we then visually inspected each eye convergence and manually set its beginning and end.

Annotation of hunting outcomes

We manually annotated the start, end and outcome of each hunting attempt. Each hunting sequence starts with eye convergence and ends when the eyes deconverge (Extended Data Fig. 1b, c). For each hunting sequence, we manually scored its outcome as success, miss, abort and reject. Success is defined as a hunting sequence that ends in the swallowing of the prey without regurgitation. Reject is defined as a hunting sequence that ends in prey ingestion followed by regurgitation. Misses and aborts are both hunting sequences that terminate with the fish deconverging the eyes without ingestion of prey. Misses were defined as the subset of cases in which the last bout of the sequence was executed towards the prey. Aborts were defined as the subset of cases where the last bout of the sequence was executed away from the prey. The hunting outcome is correlated with the last movement of a hunting sequence (Extended Data Fig. 2d–f). Successes are preferentially associated with capture swims, and misses and aborts are preferentially associated with high-angle turns and routine turns. Hunting events were annotated for 28 foraging animals (1–2 h of behavioural recording per animal in the presence of *Paramecia*), and 7 nonforaging animals (~2 h of behavioural recording per animal in the absence of *Paramecia*).

We also manually annotated suction events, which are not associated with a discrete movement bout. Suction events were observed in a small number of animals (5 out of 25) (Extended Data Fig. 2t).

HMM

We fit a Gaussian mixture model to the dispersal distance distribution and determined the number of component distributions in the behaviour by the Bayesian information criterion. The resulting label assignments were used as input to infer the underlying states of a HMM using the Baum Welch algorithm[35], implemented in Julia. Exploitation states have a significantly lower movement rate than exploration states. In Extended Data Fig. 2u, we show further subdivision of the exploitation state into a state with very low movement rate (< 0.1 Hz) and very low spatial dispersal and a state with moderate movement rate and spatial dispersal by fitting a three-state HMM to a subset of animals ($n = 16$). We find that the inverse relationship between movement rate and hunting probability is consistent regardless of whether we fit 2 or 3 states. The lower the movement rate, the more strongly individual movements are biased towards hunting. We further applied PCA to a comprehensive set of ten behavioural parameters (Extended Data Fig. 3d–f). The ten behavioural parameters included the rate of each movement type, as well as the rate of hunting initiations and outcomes. Because the HMM states were based on dispersal distance, we deliberately excluded dispersal distance from this set of ten behavioural parameters, to enable an independent check on the HMM states.

Multivariate linear regression and cross-correlation analysis

To uncover the neurons with activity related to behavioural states and individual behavioural events, we used a combination of linear regression and cross-correlation. We used cross-correlation analysis to determine the timing of neural signals with respect to individual events (for example, prey ingestion, routine turns and so on). For each class of behavioural event, a neuron must have its event-aligned average activation peak within 2 s of the event to be included in the functional network for that event class. We then use linear regression to estimate the strength of the relationship between the neural activity and each class of behavioural event. For each class of behavioural event, neurons must have a significant correlation to events of that class ($P < 0.001$) to be included in the functional network for the event class. To uncover the visual network for prey detection, the visual field was subdivided into eight angular segments (−120° to 120°). Distance-weighted *Paramecia* position within each segment was used to generate one of eight regressors for prey angles (for example, the retinotopic map of prey position in the brain in Extended Data Fig. 5f, i). *Paramecia* distance from the head of the fish was exponentially weighted ($k = -0.973\ \mathrm{mm}^{-1}$).

To verify the hypothesis that the exploitation-state-encoding neurons encoded the exploitation state and not simply hunting events, we excluded increasingly large time intervals (0.5 s to 10 s with a step of 0.5 s) encompassing all hunting events across the entire experiment (Extended Data Fig. 8a). We then recalculated the correlations between the activity of these neurons and the exploitation state after the temporal exclusion. As shown in Extended Data Fig. 8b, the dorsal raphe exploitation-state neuron population showed higher correlation than the set of all exploitation-state neurons, and the correlation coefficients of both groups were highly stable across all temporal exclusion windows from 0.5 s to 10 s. This indicates that the correlation between behavioural exploitation state and the activity of exploitation-state neurons is not simply an artefact of increased hunting-related activity during the exploitation state but can be also decoded during inter-movement intervals.

The consistently higher rate of locomotion during exploration is not sufficiently resolved by the calcium reporter to permit the removal of time intervals containing individual movement events during exploration.

PCA of whole-brain neural activity

PCA of whole-brain neural activity provided an unsupervised perspective for demonstrating the existence of exploration and exploitation states in the brain activity. We used the NMF-factored neural activity as the input to PCA. To eliminate animal handling effects at the beginning of the experiment, the first ten minutes of neural activity were excluded from PCA for each recording session. We applied PCA to the NMF-factored neural activity (without preselection, as shown in the first column of Extended Data Fig. 6) and projected the high-dimensional neural population activity into a three-dimensional PCA space for visualization (as shown in Fig. 2d, Extended Data Fig. 4b and the second column of Extended Data Fig. 6). As shown in the fourth column of Extended Data Fig. 6, all PCA trajectories were colour-coded by the normalized amplitude (ranging from 0 to 1) of the mean activity of exploitation-state neurons (third column of Extended Data Fig. 6).

k-means clustering to independently uncover exploitation-state-encoding subpopulations in the dorsal raphe

In addition to identifying exploitation-state-encoding dorsal raphe neurons by using linear regression to behaviour states, we conducted an independent clustering-based procedure to verify the existence of exploitation-state-encoding neurons in the dorsal raphe (Extended Data Fig. 8c–f). We correlated neural activity between all pairs of dorsal raphe neurons to generate a correlation matrix for each fish. An unsupervised *k*-means clustering analysis was then applied to the pairwise correlation matrix to cluster dorsal raphe neurons into five clusters. We selected the tightest-correlated cluster (with highest average intracluster correlation) from each animal and examined its neural activity during the transition of exploration states to exploitation states. In every fish (17 out of 17), the tightest-correlated dorsal raphe cluster showed significantly higher activity in exploitation state than exploration state.

Comparison of correlation within and across fish

After selecting the tightest-correlated cluster in each animal, we calculated the correlation C_{ij}^{k} between the neural activity of neuron k in the dorsal raphe subpopulation of fish i, and the behavioural state of fish j. We then calculated the mean correlation coefficient C_{ij} between fish i and fish j by averaging over each dorsal raphe neuron k in the selected cluster. As shown in Extended Data Fig. 8e, the neural activity of the dorsal raphe subpopulation is reliably correlated with the behaviour state of the same fish but not with the behavioural state of other fish, which confirms that these are real correlations and not spurious. For quantitative comparison, we compared C_{ij}^{k} under two conditions: $i=j$ (on-diagonal blocks of Extended Data Fig. 8e) and $i \neq j$ (off-diagonal blocks of Extended Data Fig. 8e). We observed that the mean of C_{ij}^{k} for $i=j$ was 0.468, and the mean of C_{ij}^{k} for $i \neq j$ was 0.024. This difference was highly significant (Extended Data Fig. 8f, $P < 10^{-8}$, two-sample *t*-test).

In brain regions that do not encode exploitation state such as the inferior olive (Extended Data Fig. 8g–j), the cluster with the highest average intracluster correlation is not persistently active during the exploitation state (Extended Data Fig. 8g, h). Furthermore, repeating the correlation analysis within and across animals, the difference is not significant (Extended Data Fig. 8i, j). The mean correlation between behavioural state and dorsal raphe state neuron activity in the same animal was −0.047, whereas the correlation between the behavioural state of one animal and the exploitation state activity of another was 0.023 (Extended Data Fig. 8j, $P=1$, two-sample *t*-test).

Normalization of dorsal raphe state activity

To facilitate pooling of data across animals in Fig. 3, dorsal raphe state activity was normalized for each animal to a range of 0 to 1 according to the formula

$$X_N = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

In Fig. 3a–d, $X_{\min}$ and $X_{\max}$ are defined as the mean of the bottom and top tenth percentile, respectively, of dorsal raphe state activity across the entire experiment. Normalization was applied to 17 animals. One of the 17 animals was excluded for analysis owing to the low number of total successful hunting attempts. In Fig. 3e–k, we compute the correlation between normalized dorsal raphe state activity and activity of sensorimotor neurons throughout the brain at the time of hunting initiation. For each animal, we generate a vector of dorsal raphe state activity at the time of hunting initiation. This vector is then normalized to range from 0 to 1 by defining $X_{\min}$ and $X_{\max}$ to be the minimum and maximum values in the vector, respectively.

Spatial *P*-value analysis

We developed a spatial *P*-value analysis method to evaluate the significance of spatial colocalization of functionally classified neurons across animals. In brief, we registered all 17 functionally imaged brains to a common reference zebrafish brain from the Z-Brain atlas[23], computed the distance of each neuron to its nearest neighbour (sharing the same state association) across all coregistered brains and then compared to a null model defined by random sampling. As shown in Extended Data Fig. 5a, the spatial *P*-value analysis followed a two-step procedure. In step 1, we applied ANTs registration to transform the live fluorescent brain volumes of 17 animals into a common reference space defined by Z-Brain. In step 2, we calculated the spatial *P* value for each functionally classified neuron in each animal by examining its spatial colocalization across all 17 fish (Extended Data Fig. 5a). For instance, to calculate the spatial *P* value of a functionally classified neuron in fish i (shown as red star), we calculate the shortest distance dr to any neuron in fish j that was classified into the same functional class to generate a same-class distance vector $[dr_{i,1}, \ldots, dr_{i,j-1}, dr_{i,j}, \ldots, dr_{i,m}]$, in which $m=17$. We then calculated the shortest distance dn between the neuron in fish i and a random size-matched population of neurons in fish j obtained by randomly sampling all neurons in fish j, with the number of neurons sampled equal to the number of neurons in the functional class in fish j. Therefore, a random distance vector $[dn_{i,1}, \ldots, dn_{i,j-1}, dn_{i,j}, \ldots, dn_{i,m}]$ was calculated. Before performing a scalar statistical test to obtain the spatial *P* value, we adopted a projection method to handle the possibility of 1 or 2 outlier fish. The projection function used in our analysis was the mean of the lowest 90% of the distances within each distance vector, resulting in scalar distances dr and dn. We then repeated the random sampling process N times (N=10,000 in our analysis) to construct a null distribution of the shortest distances between a functionally classified neuron in fish i and the randomly sampled neurons in other fish. We then fit this null distribution with a normal distribution. The *P* value for this functionally classified neuron in fish i relative to all other fish was then calculated by comparing dr to the normally distributed null model obtained from sampling dn. We only reject the null hypothesis if a functionally classified neuron in fish i has a spatial *P* value < 0.025—that is, if the probability that the null model generating a distance less than or equal to the projected distance is less than 0.025.

Neuromodulation of sensorimotor activity by dorsal raphe state neurons

We analysed the activity of sensory and premotor neurons throughout the brain as a function of dorsal raphe state activity by evaluating the significance of their correlation. To test the statistical significance of the correlation value for each neuron, we created a null distribution of correlation values by randomly permuting the activity of each neuron 10,000 times. For each neuron, we compared the null distribution of

its correlation values against its original correlation value (neurons with P value < 0.05 are shown in Fig. 3e, f).

A stochastic nonlinear oscillator model of internal state

We developed a stochastic nonlinear dynamical system model of exploration and exploitation (Extended Data Fig. 7e, f):

$$\frac{dT}{dt} = \sum_i A_i^{spont}\delta(t - t_i^{spont}) + \sum_i A_i^{evoked}\delta(t - t_i^{evoked}) - \gamma T$$
$$\frac{dS}{dt} = \alpha T - \beta \cdot 1_{S>0}$$

in which

$$1_{S>0} \equiv \begin{cases} 1, S > 0 \\ 0, S = 0 \end{cases}$$

is an indicator function for exploitation state.

When the state variable is low ($S = 0$), the system is in exploration state. Spontaneous trigger events arrive according to a Poisson process with rate λ_0. The duration D_0 of the exploration state is exponentially distributed, $D_0 \sim \exp(\lambda_0)$, with an average waiting time of $1/\lambda_0$. Each spontaneous trigger event is represented as a delta function $\delta(t - t_i^{spont})$, representing a spontaneous event at time t_i^{spont}, multiplied by a stochastic amplitude A_i^{spont}. We model both the amplitude A_i^{spont} and the waiting time between spontaneous trigger events $T_i^{spont} \equiv t_i^{spont} - t_{i-1}^{spont}$ as exponentially distributed random variables, with $A_i^{spont} \sim \exp(\lambda_1)$ and $T_i^{spont} \sim \exp(\lambda_0)$. With each spontaneous event, the trigger signal rises impulsively to A_i^{spont}, and then decays exponentially at rate γ. We can configure the model parameters such that the state variable $S(t)$ is rapidly driven to $k A_i^{spont}$ (with k determined primarily by α and γ), and then undergoes slow linear decay at rate β. While the state variable is elevated ($S > 0$), the system is in exploitation state. The duration D_i^{spont} of the exploitation state is set by $k A_i^{spont}/\beta$, so duration D_i^{spont} follows an exponential distribution, $D_i^{spont} \sim \exp(\lambda_1 \beta / k)$. When fitting the model to empirical data, we fit the event amplitudes A_i^{spont} and times t_i^{spont} individually for each transition. The remaining model parameters (α, β and γ) were obtained by simultaneously fitting the model equations for $T(t)$ and $S(t)$ to the dynamics of the activity of trigger and exploitation state neurons during 14 exploitation states having a duration of at least 10 min (14 out of 17 fish had state durations of ≥10 min). The joint fitting procedure yielded $\alpha = 1.5$, $\beta = 0.1\ min^{-1}$ and $\gamma = 1.3\ min^{-1}$.

Immunostaining

We first performed whole-brain calcium imaging during foraging for 80 min, and then immediately fixed and immunostained the brain for neuromodulatory markers such as serotonin[36–38], dopamine[22,38] or acetylcholine[39] using the MultiMAP method[20,21] to relate antibody-positive neurons to their neural activity. The immunohistochemical method to fluorescently label larval zebrafish neural populations used for these experiments, as well as at least ten control embryos for each antibody, is a slightly modified version of the MultiMAP method published previously[20]. We found this method yielded the same neuronal labelling patterns as both the standard PFA immunolabelling methods and published gene expression patterns[40]. In contrast to the standard PFA immunolabelling technique, our modified protocol yielded better antibody penetration throughout the entire brain. As previously described[20], we found that by embedding the intact fish, the integrity of brain structures was preserved, allowing registration of live neurons with immunolabelled neurons.

In brief, after functional imaging, larvae were embedded in blocks of 2.5% intermediate melt agarose, immersed in ice-cold 4% PFA in phosphate buffered saline with 0.25% Triton-X100 (PBST) and incubated on a rocker at 4 °C overnight.

Larvae were rinsed 3 times in ice-cold PBST buffer, then incubated for several hours with PBST buffer, changed every 2 h at 4 °C. All antibody incubations were performed at 4 °C. Primary antibodies were added at a 1:500 dilution in PBST and rocked for 72 h, followed by several hours of washing with PBST buffer, changed every 2 h. Secondary antibodies were added at a 1:500 dilution in PBST and rocked for another 72 h, followed by several hours of washing with PBST. To image, embedded fish were mounted in 2% agarose and imaged using a Zeiss 880 confocal microscope with a water immersion 20× objective. Z stacks were collected with pixel dimensions of 0.998 µm × 0.998 µm × 2 µm. The image size was 638.9 µm × 638.9 µm.

The primary antibody used to label dopamine neurons was mouse anti-tyrosine hydroxylase (Immunostar 22941). The primary antibody used to label serotonin neurons was rabbit anti-5HT (Immunostar 20080). The primary antibody used to label acetylcholine neurons was goat anti-choline acetyltransferase (Millipore AB144P).

The secondary antibodies used were donkey anti-mouse Alexa647 (Jackson ImmunoResearch 715-606-151), donkey anti-rabbit Alexa594 (Jackson ImmunoResearch 711-586-152) and donkey anti-goat Alexa647 (Jackson ImmunoResearch 705-606-147).

For display purposes only, Supplementary Video 4 of dopaminergic and acetylcholinergic brain stacks were gamma-adjusted with a gamma value of 1.7 in Photoshop.

Satiety control experiment and starved control experiment

We first examined the relation between exploitation/exploration transitions and changes in satiety by placing larvae in a behavioural arena without prey, which has no change in satiety from prey consumption. As expected, the probability of hunting-related movements is lower in the absence of prey (Extended Data Fig. 4d). However, all animals continue to show state-dependent modulation of locomotion and hunting-related movements in the absence of prey (Extended Data Fig. 4d). Animals continue to prefer routine turns and slow swims 2 during exploration and hunting-related movements such as J-turns during exploitation (Extended Data Fig. 4d). Overall the amount of time spent in the exploitation state is slightly elevated in the absence of prey, but this difference was not statistically significant (Extended Data Fig. 4c).

We further tested whether satiety can act as a modulator of exploitation and exploration state dynamics. We compared larvae that were fed until the start of the experiment with siblings that were starved for 24 h. We observed that starved fish also showed an increase in the probability of hunting-related movement such as J-turns (Extended Data Fig. 4f) across the entire experiment, suggesting that hunger can also globally elevate hunting. Thus, an intrinsic mechanism exists to drive spontaneous oscillations in decision-making during foraging.

Animal care and transgenic lines

All animal experiments were approved by the Harvard University Institutional Animal Care and Use Committee. All imaging experiments used Tg(elavl3:H2B-GCaMP6s$^{+/+}$)[15] with nacre (mitfa$^{-/-}$). Behavioural experiments used Tg(elavl3:H2B-GCaMP6s$^{+/+}$) and Tg(elavl3:GCaMP6s$^{+/+}$)[9], both with nacre (mitfa$^{-/-}$). All zebrafish (*Danio rerio*) were 6–8 dpf. Fish were reared on a 12:12 h light:dark cycle at 25 °C. They were maintained in groups of 20 in E3 water and were fed *Paramecia* daily.

Statistical analysis

Sample sizes and statistics are reported in the figure legends and text for each measurement. For comparison of behaviour across exploration and exploitation states, we used a two-tailed Wilcoxon signed-rank test. Adjusted P values are corrected for multiple comparisons by the Holm–Bonferroni method. Box plots depict the median (horizontal line), first and third quartiles (box), and the smallest and largest data within 1.5-fold of the interquartile range of the first and third quartiles, respectively. Sample size was not predetermined by a statistical method. The experiments were not randomized and were not performed with blinding.

Computer hardware for data acquisition and analysis

The tracking microscope was controlled by a rack-mounted eight-core (i7-5960X, Intel) workstation with two GPUs (GTX 1080, NVIDIA). Offline data analysis was performed using four rack-mounted eight-core Linux servers (i7-6900K, Intel) with four GPUs each (GTX 1080 Ti, NVIDIA).

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The data supporting the findings of this study are available from the corresponding authors upon request.

Code availability

All software was written in Julia 0.6.3 and C++11 using the Julia package ecosystem (Optim.jl, PyPlot.jl, Images.jl, Cairo.jl, HDF5.jl, Lasso.jl, CUDAdrv.jl and CUDArt.jl). GPU processing was implemented for fish tracking, as well as online and offline image registration. Image registration and data analysis code was based on published work[9,10,16,21,32,33], and is available upon request.

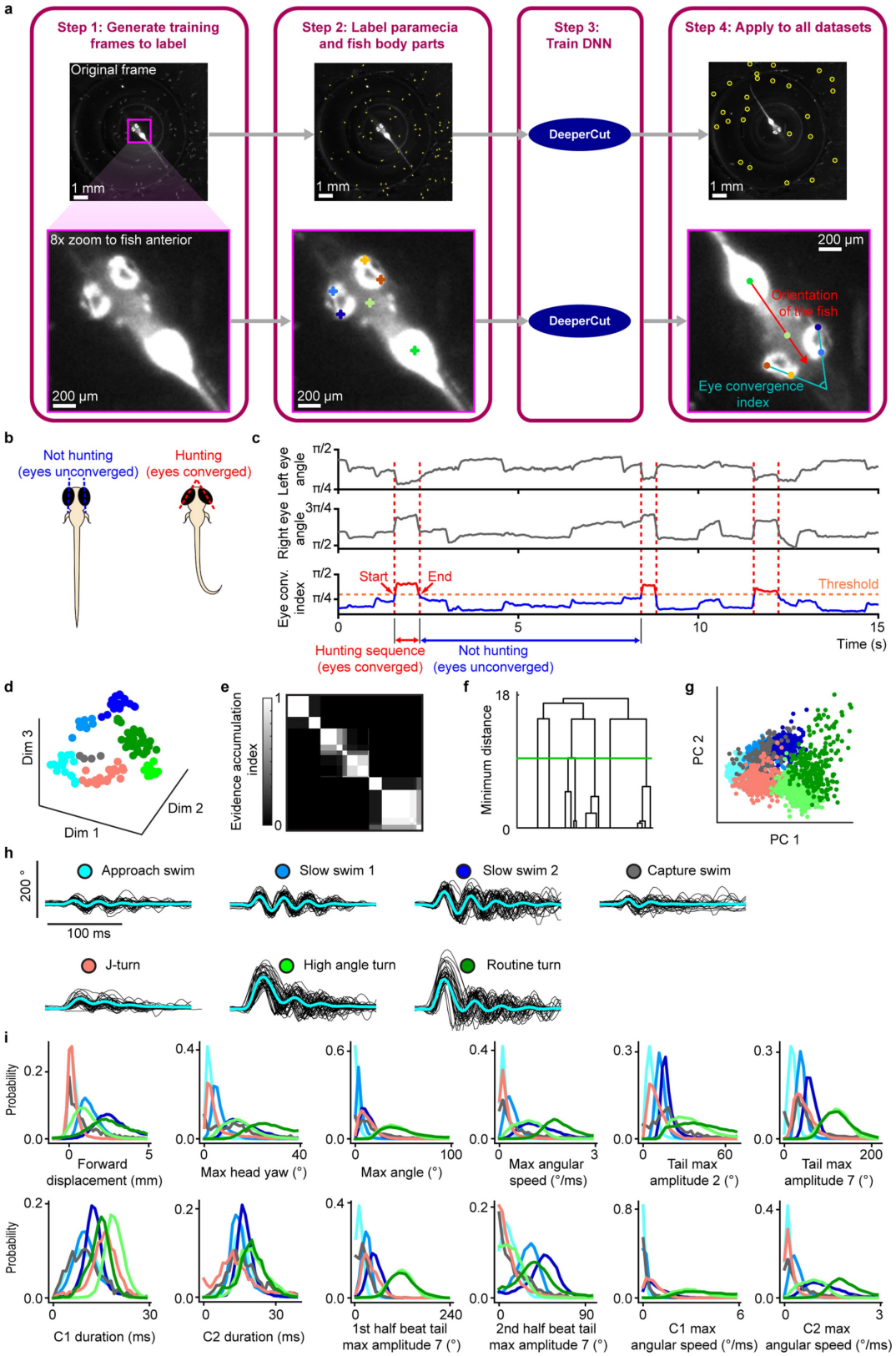
29. Avants, B., Tustison, N. & Song, G. Advanced Normalization Tools (ANTS). *Insight J.* **2**, 1–35 (2009).
30. van der Maaten, L., Hinton, G. E., van der Maaten, L. & Hinton, G. E. Visualizing high-dimensional data using *t*-SNE. *J. Mach. Learn. Res.* **9**, 2579–2605 (2008).
31. Sibson, R. SLINK: An optimally efficient algorithm for the single-link cluster method. *Comput. J.* **16**, 30–34 (1973).
32. Insafutdinov, E., Pishchulin, L., Andres, B., Andriluka, M. & Schiele, B. in *European Conference on Computer Vision* 34–50 (Springer, 2016).
33. Mathis, A. et al. DeepLabCut: markerless pose estimation of user-defined body parts with deep learning. *Nat. Neurosci.* **21**, 1281–1289 (2018).
34. Nath, T. et al. Using DeepLabCut for 3D markerless pose estimation across species and behaviors. *Nat. Protoc.* **14**, 2152–2176 (2019).
35. Rabiner, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE* **77**, 257–286 (1989).
36. Lillesaar, C., Tannhäuser, B., Stigloher, C., Kremmer, E. & Bally-Cuif, L. The serotonergic phenotype is acquired by converging genetic mechanisms within the zebrafish central nervous system. *Dev. Dyn.* **236**, 1072–1084 (2007).
37. Lillesaar, C., Stigloher, C., Tannhäuser, B., Wullimann, M. F. & Bally-Cuif, L. Axonal projections originating from raphe serotonergic neurons in the developing and adult zebrafish, Danio rerio, using transgenics to visualize raphe-specific pet1 expression. *J. Comp. Neurol.* **512**, 158–182 (2009).
38. McLean, D. L. & Fetcho, J. R. Ontogeny and innervation patterns of dopaminergic, noradrenergic, and serotonergic neurons in larval zebrafish. *J. Comp. Neurol.* **480**, 38–56 (2004).
39. Hong, E. et al. Cholinergic left-right asymmetry in the habenulo-interpeduncular pathway. *Proc. Natl Acad. Sci. USA* **110**, 21171–21176 (2013).
40. Panula, P. et al. The comparative neuroanatomy and neurochemistry of zebrafish CNS systems of relevance to human neuropsychiatric diseases. *Neurobiol. Dis.* **40**, 46–57 (2010).

Acknowledgements We thank M. Burns, M. Frank, S. Flavell, P. Dayan, K. Taute and G. Rainer for discussions and feedback; A. Mathis and M. W. Mathis for providing unpublished code, help with the implementation of DeepLabCut for tracking *Paramecia* and zebrafish, and for motivating us to use deep learning methods; and we thank M. Ahrens and H. Baier for Tg(elavl3:H2B-GCaMP6s) and I. Bianco for discussions about prey capture. This work was financially supported by the Rowland Institute at Harvard.

Author contributions J.M.L and D.N.R. conceived the project on state-dependent modulation of foraging behaviour, identified behavioural states (for example, using HMM and PCA) and state-related and hunting-related neural populations, created the dynamical system model of motivational states and provided overall guidance on analysis of behaviour and neural activity. J.C.M. contributed to the conception of whole-brain imaging of hunting sequences using the tracking microscope, conducted all experiments, contributed to experimental design, classified individual movements into bout types, annotated and confirmed each hunting sequence and outcome, analysed state-dependent modulation of behaviour and contributed to identification and functional analysis of neuromodulatory populations. M.L. trained deep neural networks for annotation of *Paramecia* and fish body parts, conducted PCA of whole-brain neural activity, analysed the functional role of neuromodulatory populations (such as by registration of live and immunostained brain volumes) and identified state-related and hunting-related neural populations across animals (spatial *P* value analysis and *k*-means clustering). D.S. designed and performed all immunostaining experiments. J.M.L., D.N.R., J.C.M. and M.L. wrote the manuscript.

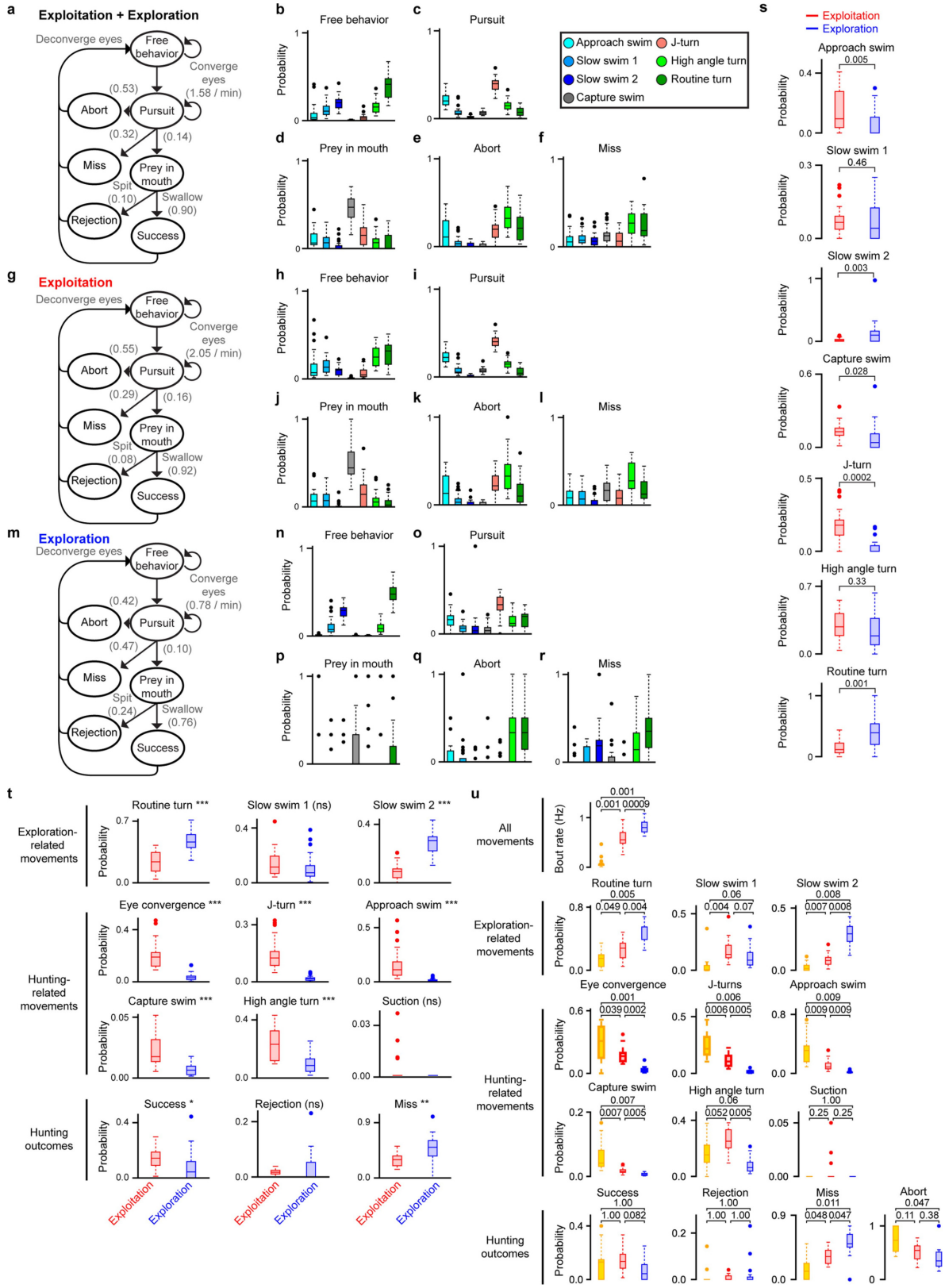
Competing interests The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1858-z.
Correspondence and requests for materials should be addressed to D.N.R. or J.M.L.
Peer review information *Nature* thanks Mario De Bono, Yonatan Loewenstein, Ethan Scott and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



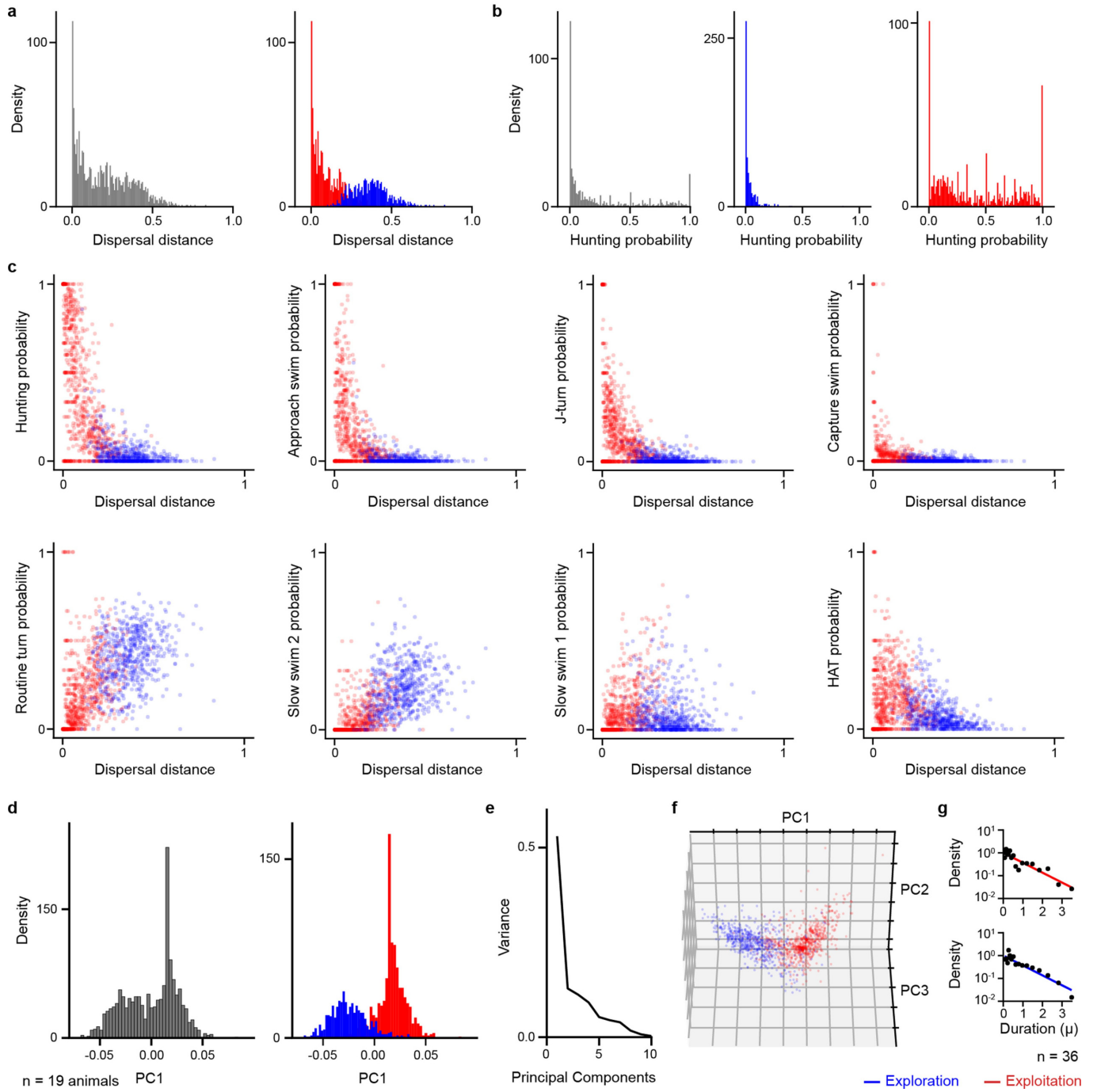
Extended Data Fig. 1 | See next page for caption.

Extended Data Fig. 1 | DNN pipeline for automated tracking of *Paramecia* and fish and density valley clustering identifies seven bout types performed by larvae during foraging. a, Automated tracking of *Paramecia* and fish followed a four-step procedure (Methods). The rightmost panel shows the automated annotation of *Paramecia* position (yellow circles), eye position (dark blue, blue, yellow and red dots), yolk position (green dot) and brain position (light green dot). Fish heading (red line) and eye convergence index (angle in cyan) were calculated using eye, and yolk, and brain positions (Methods). **b**, Schematic of eye convergence index when the fish is not hunting (left) or hunting (right). **c**, Calculation of the eye convergence index. Top, angles of the left and right eyes are calculated from the two detected points of each eye. Bottom, eye convergence index is calculated as the difference of the individual eye angles. By setting a threshold for the eye convergence index, we define the start and end of each hunting sequence. **d**–**i**, The bout clusters that larvae executed during functional imaging with the tracking microscope (25 larvae in the presence of *Paramecia* and 7 in the absence of *Paramecia*) were identified by clusterdv (Methods). To estimate the similarity between bout clusters, pairwise ROC curves were calculated. **d**, Dimensionality reduction using *t*-distributed stochastic neighbour embedding (*t*-SNE). **e**, Evidence accumulation matrix of clusterdv assignments was constructed by repeating the *t*-SNE embedding 100 times. The rows and columns of the matrix were sorted by single link clustering. **f**, Dendrogram of **e**. Green line marks the largest jump on the dendrogram and determines the number of bouts types (seven) that were detected. **g**, First 2 PCs of 73 kinematic parameters that were calculated for each bout. Colours represent bout categories from **f**. **h**, Angle of caudal tail segment versus time of 50 randomly picked bouts of each type. Cyan lines are the average of all bouts in each category. **i**, Bout kinematic distributions were obtained by using all the bouts of each type. Colours represent bout types as in **h**.

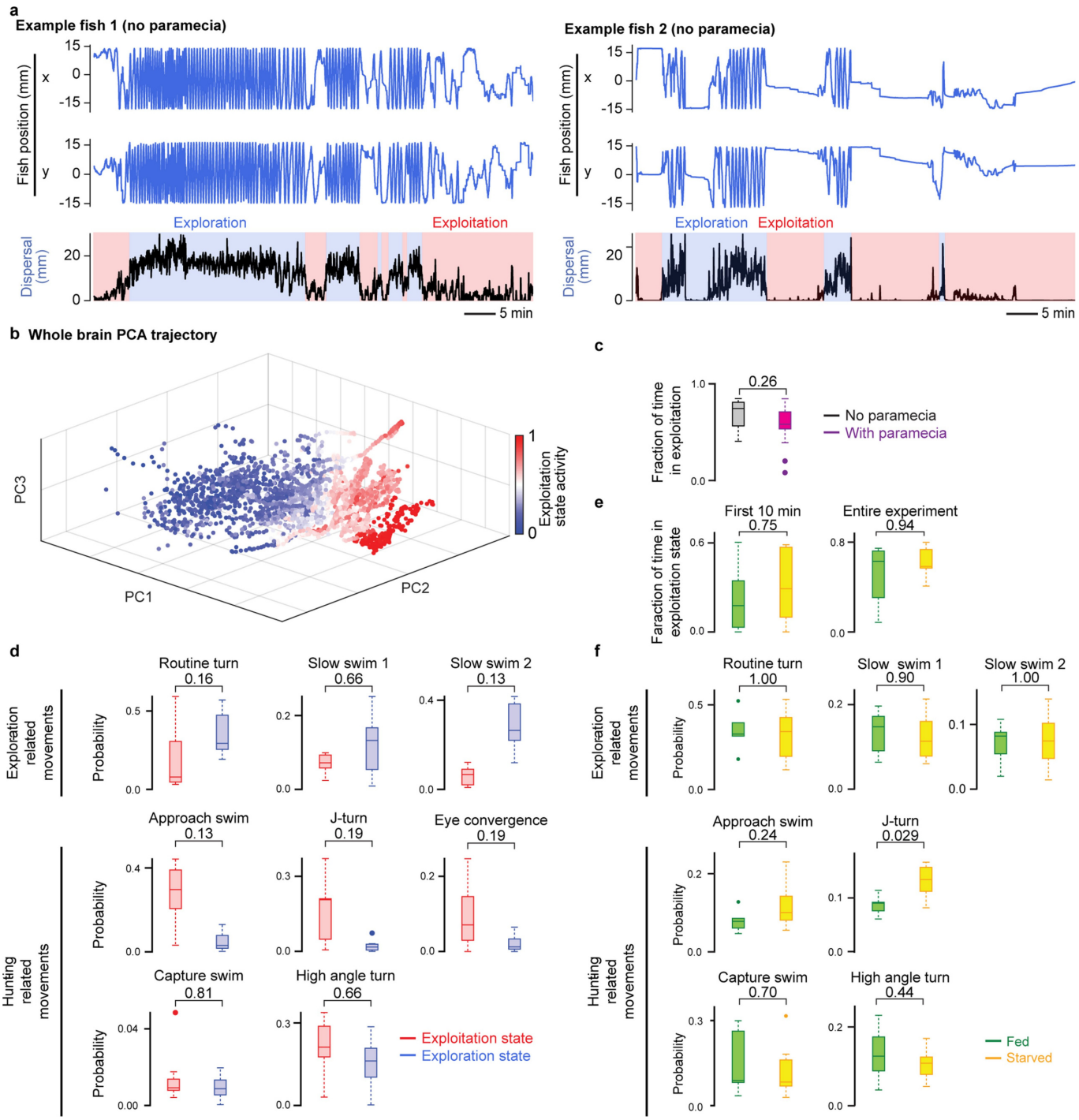


Extended Data Fig. 2 | See next page for caption.

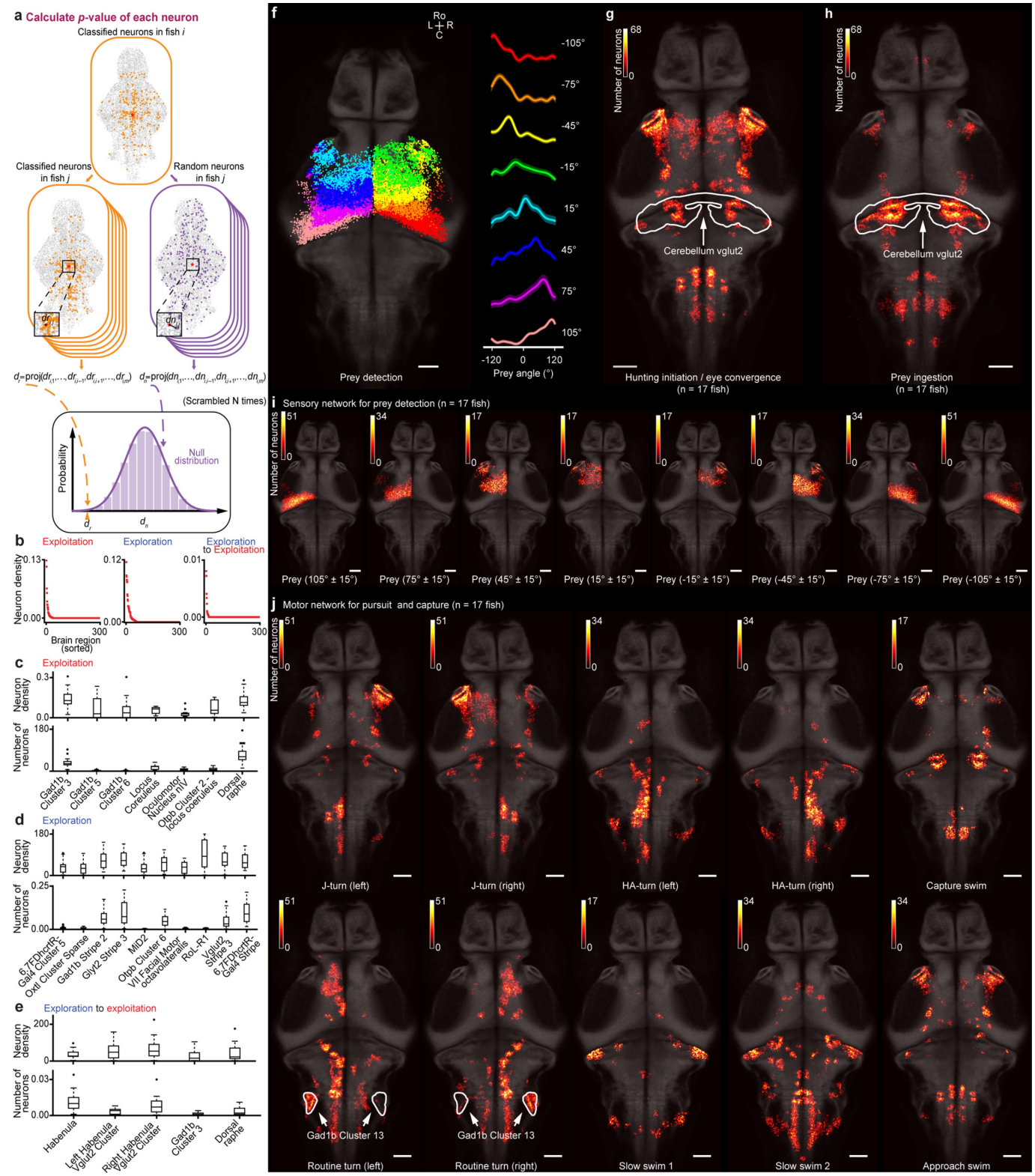
Extended Data Fig. 2 | Sequential Markov decision process of larval foraging behaviour, final hunting bout is modulated by foraging state, and movement rate and hunting probability are inversely related. a–**r**, Markov decision process of larval foraging behaviour. A schematic of foraging behaviour as a Markov decision process is shown for all time (**a**), exploitation state (**g**) or exploration state (**m**). Each node represents a decision point in the hunting sequence, and each arrow has a state-dependent transition probability (grey number next to arrow). The 'free behaviour' and 'pursuit' nodes are composed of sequences of multiple bout types that may occur with distinct probabilities. **b**, **c**, **h**, **i**, **n**, **o**, Bout type probabilities for the free behaviour node (**b**, **h**, **n**), which occurs when the eyes are unconverged, and the pursuit node (**c**, **i**, **o**), which takes place when the eyes are converged and the fish is chasing prey. **d**–**f**, **j**–**l**, **p**–**r**, Bout type probabilities of the last bout of the pursuit sequence for each of the hunting outcomes: prey in mouth (**d**, **j**, **p**), abort (**e**, **k**, **q**) or miss (**f**, **l**, **r**). Once the prey is in the mouth, the fish can swallow (success) or spit it out (rejection). The colours in the box plots represent each observed bout type (legend, top right corner). Statistical analysis is presented in Supplementary Table 1. **s**, Movement type that the fish executes at the end of each hunting sequence during exploitation and exploration ($n = 25$ animals, 2,972 hunting attempts). **t**, Box plots of state-dependent differences in the probability of exploration-related movement types (top), probability of hunting-related movement types (middle) and probability of hunting outcomes (bottom). Probability of a movement type is defined as the fraction of all movements that are of a given type, and probability of hunting outcome is defined as the fraction of all hunting sequences that end with a given outcome. Adjusted P values for state-dependent differences: routine turn (7.37×10^{-5}), slow swim type 1 (0.092), slow swim type 2 (9.83×10^{-5}), eye convergence (6.14×10^{-5}), J-turn (8.60×10^{-5}), approach swim (9.83×10^{-5}), capture swim (7.63×10^{-5}), high angle turn (1.0×10^{-5}), suction (0.25), success (0.020), rejection (0.63) and miss (3.5×10^{-3}). $n = 25$ animals, 2,972 hunting attempts and 88,814 movements. **u**, To more finely parse state-dependent behaviour, we further subdivide the dispersal data by fitting a 3-state HMM. This results in a state with very low movement rate (< 0.1 Hz, yellow) and spatial dispersal, a state with intermediate movement rate and spatial dispersal (red) and a state with very high movement rate and dispersal (blue). We find that the inverse relationship between movement rate and hunting probability is consistent regardless of whether we fit two (**t**) or three states (**u**). The lower the movement rate, the more strongly individual movements are biased towards hunting ($n = 16$ animals, 1,605 hunting attempts and 40,069 movements). **s**–**u**, Pairwise comparisons between the states was performed by using a two-tailed Wilcoxon signed-rank test. Adjusted P values have been corrected for multiple comparisons by Holm–Bonferroni method. $^{*}P < 0.05$, $^{**}P < 0.01$ and $^{***}P < 0.001$.



Extended Data Fig. 3 | Distribution of behavioural parameters across exploitation and exploration state. a, Distribution of dispersal distance (left), and colour-coded (right) by whether the observation occurred during exploration (blue) or exploitation (red) state, as classified by the HMM (Methods). To facilitate comparison across animals, the dispersal distance was normalized to the range of each animal. **b**, Distribution of hunting probability across all time (left), during exploration state (blue) and during exploitation state (red). **c**, Distribution of probability for hunting, approach swim, J-turn, capture swim, routine turn, slow swim type I, slow swim type II and high-angle turn, plotted against dispersal distance, and colour-coded by whether the observation occurred during exploitation (red) or exploration (blue) state. **d**–**f**, PCA of foraging behaviour reveals a bimodal distribution of behavioural variables. PCA was performed using ten behavioural variables: rate of each movement type (approach swim, J-turn, capture swim, routine turn, slow swim type I, slow swim type II and high-angle turn), rate of all hunting sequences, rate of successful hunting sequences and rate of aborted hunting sequences (Methods). **d**, Distribution of the first PC (left), and colour-coded (right) by whether the observation was classified as exploration (blue) or exploitation (red) by the HMM. **e**, Variance explained by each PC. **f**, Distribution of all behavioural parameters projected into the first three PC dimensions, colour-coded by whether the observation occurred during exploitation (red) or exploration (blue) state. For **a**–**f**, $n = 19$ animals. **g**, Normalized state duration distributions for exploitation (top) and exploration (bottom) ($n = 36$ animals), divided into 20 bins. The state durations of each animal were normalized by the mean state duration of that animal (normalization was separately performed for exploitation state and exploration state), and then used to fit an exponential model by maximum likelihood (red and blue lines).

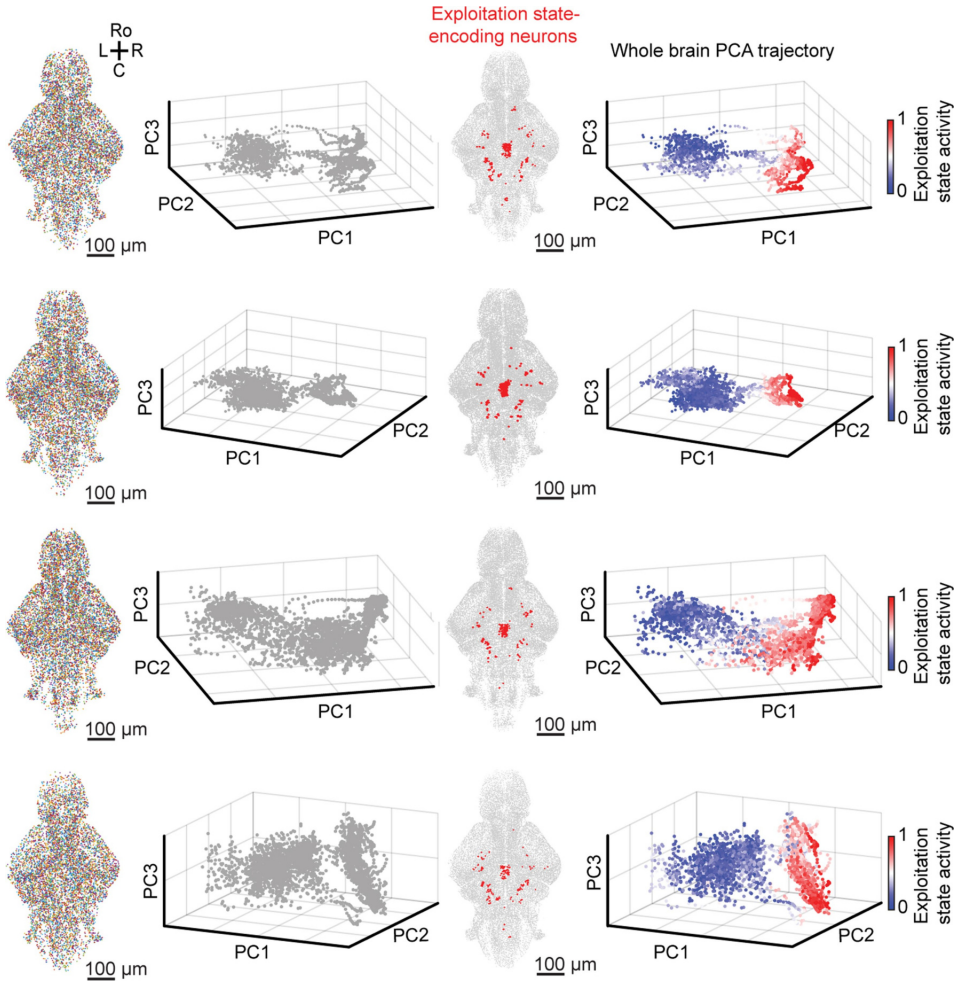


Extended Data Fig. 4 | Exploitation–exploration state transitions occur in the absence of prey and are not modulated by starvation. a, Behavioural state transitions in two animals, each placed in an arena without *Paramecia* for over an hour. **b**, PCA trajectory based on whole-brain activity from a representative animal in the absence of prey, colour-coded by the activity of exploitation-state encoding neurons. **c**, Fraction of time in exploitation state in the absence (grey) and presence (purple) of *Paramecia* (25 animals in the presence of *Paramecia* and 7 animals in the absence of *Paramecia*). **d**, Box plots of state-dependent differences in movement types ($n = 7$ animals, 13,997 movement events). The number of capture and rejection events were both 0, because there was no prey. **e**, **f**, Comparison between starved and fed larvae. **e**, Fraction of time in exploitation state for the first 10 min (left), and the entire duration of the experiment (right). **f**, Probabilities by movement type for larvae that where fed (green) or starved for 24 h before the experiment (yellow) ($n = 5$ animals in fed condition, $n = 9$ animals in 24 h starved condition). **c**–**f**, Adjusted *P* value (above each plot) obtained by two-tailed Wilcoxon signed-rank test. Adjusted *P* values have been corrected for multiple comparisons by Holm–Bonferroni method.

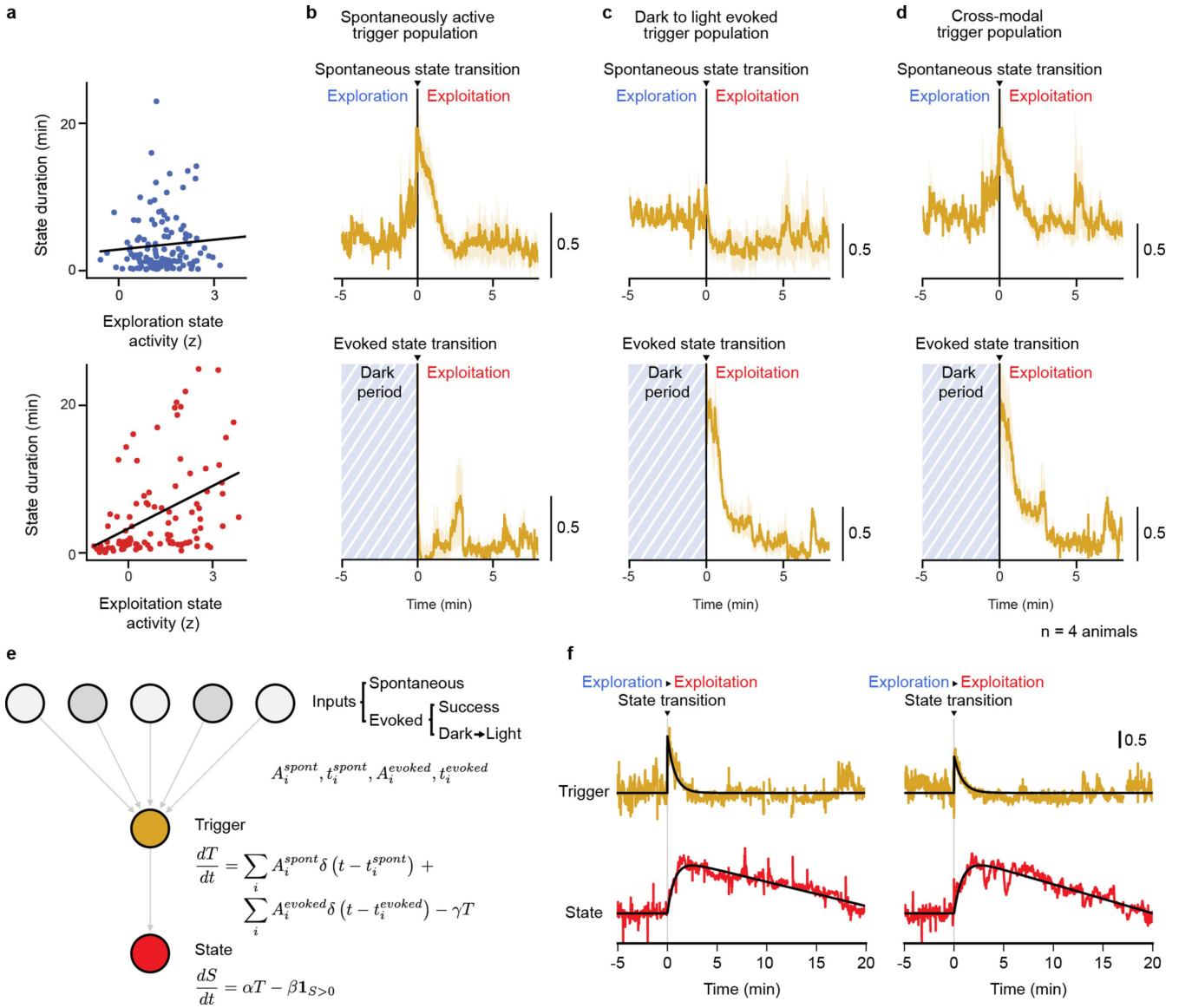


Extended Data Fig. 5 | See next page for caption.

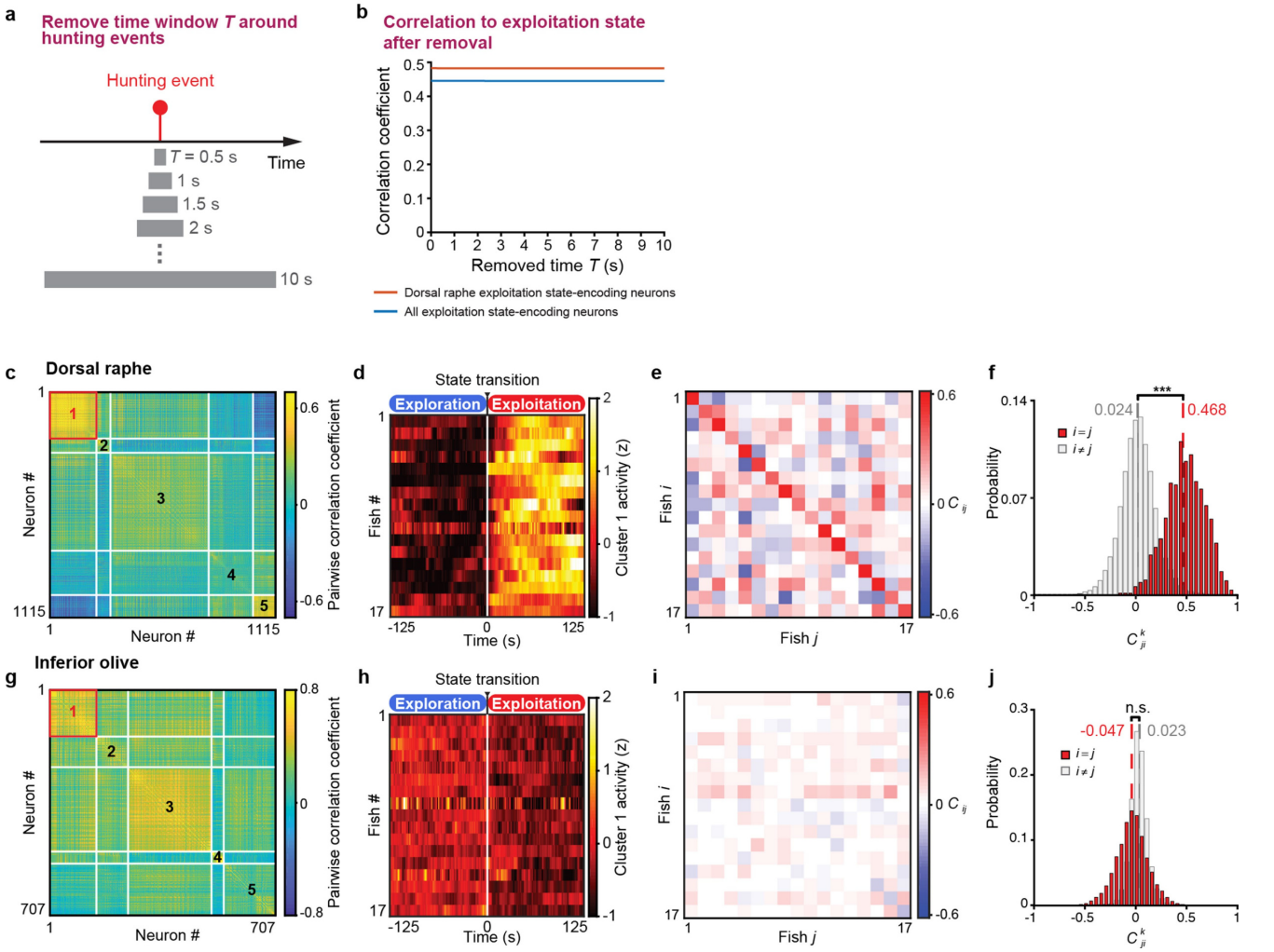
Extended Data Fig. 5 | Spatial *P* value analysis, spatial distribution of neurons encoding exploitation state, exploration state, and transition from exploration to exploitation, and sensorimotor networks activated during hunting. a, The spatial *P* value analysis followed a two-step procedure. In step 1, ANTs registration was performed to transform the live fluorescent brain volumes of 17 animals into a common reference space defined by Z-Brain. In step 2, a spatial *P* value was calculated for each functionally classified neuron in each animal by examining its spatial colocalization across all 17 fish. For instance, to calculate the spatial *P* value of a functionally classified neuron in fish *i* (shown as red star), we calculate the shortest distance *dr* to any neuron in fish *j* that was classified into the same functional class to generate a same-class distance vector $dr_{i,1},\ldots,dr_{i,j-1},dr_{i,j},\ldots,dr_{i,m}$, in which $m = 17$. We then calculated the shortest distance *dn* between the neuron in fish *i* and a random size-matched population of neurons in fish *j* obtained by randomly sampling all neurons in fish *j*, with the number of neurons sampled equal to the number of neurons in the functional class in fish *j*. Therefore, a randomized distance vector $dn_{i,1},\ldots,dn_{i,j-1},dn_{i,j},\ldots,dn_{i,m}$ was calculated. To facilitate comparison of these distance vectors, we reduce them to a single scalar value by calculating the mean of the lowest 90% of the distances within the same-class distance vector and random distance vector. We then repeated the random sampling process *N* times ($N = 10{,}000$ in our analysis) to construct a null distribution of the shortest distance between a functionally classified neuron in fish *i* and the randomly sampled neurons in other animals. We then fit this null distribution with a normal distribution. The *P* value for this functionally classified neuron in fish *i* relative to all other animals was then calculated by evaluating *dr* on the basis of a normal distribution model. We reject the null hypothesis of random sampling if a functionally classified neuron in fish *i* has a spatial *P* value < 0.025. **b**, The spatial density for each area in the Z-Brain map was calculated for neurons that encode exploitation (left), exploration (middle), and transition from exploration to exploitation (right). **c**–**e**, Brain regions with neural densities higher than two standard deviations (measured across all brain regions) for exploitation-state-encoding neurons (**c**), exploration-state-encoding neurons (**d**) and trigger neurons that encode the transition from exploration to exploitation (**e**) ($n = 17$ animals; only neurons with consistent spatial localization and spatial *P* value < 0.025 are considered for this analysis). **f**, Retinotopic map of prey detection neurons in the optic tectum. Left, anatomical position of neurons tuned to different prey angles (Methods). Right, angular tuning curves of optic tectum neurons classified by angle of peak prey response, in 30° increments. **g**, Location of all neurons time-locked to the onset of eye convergence. **h**, All neurons time-locked to successful prey ingestion. **i**, Retinotopic map for each prey detection angle depicted individually across eight maps. **j**, Motor networks that drive distinct movement types associated with pursuit and capture of prey. Motor networks were identified for all seven observed movement types. All panels depict neurons with spatially significant colocalization across animals (spatial *P* value < 0.025, $n = 17$ animals). Scale bars, 50 µm.



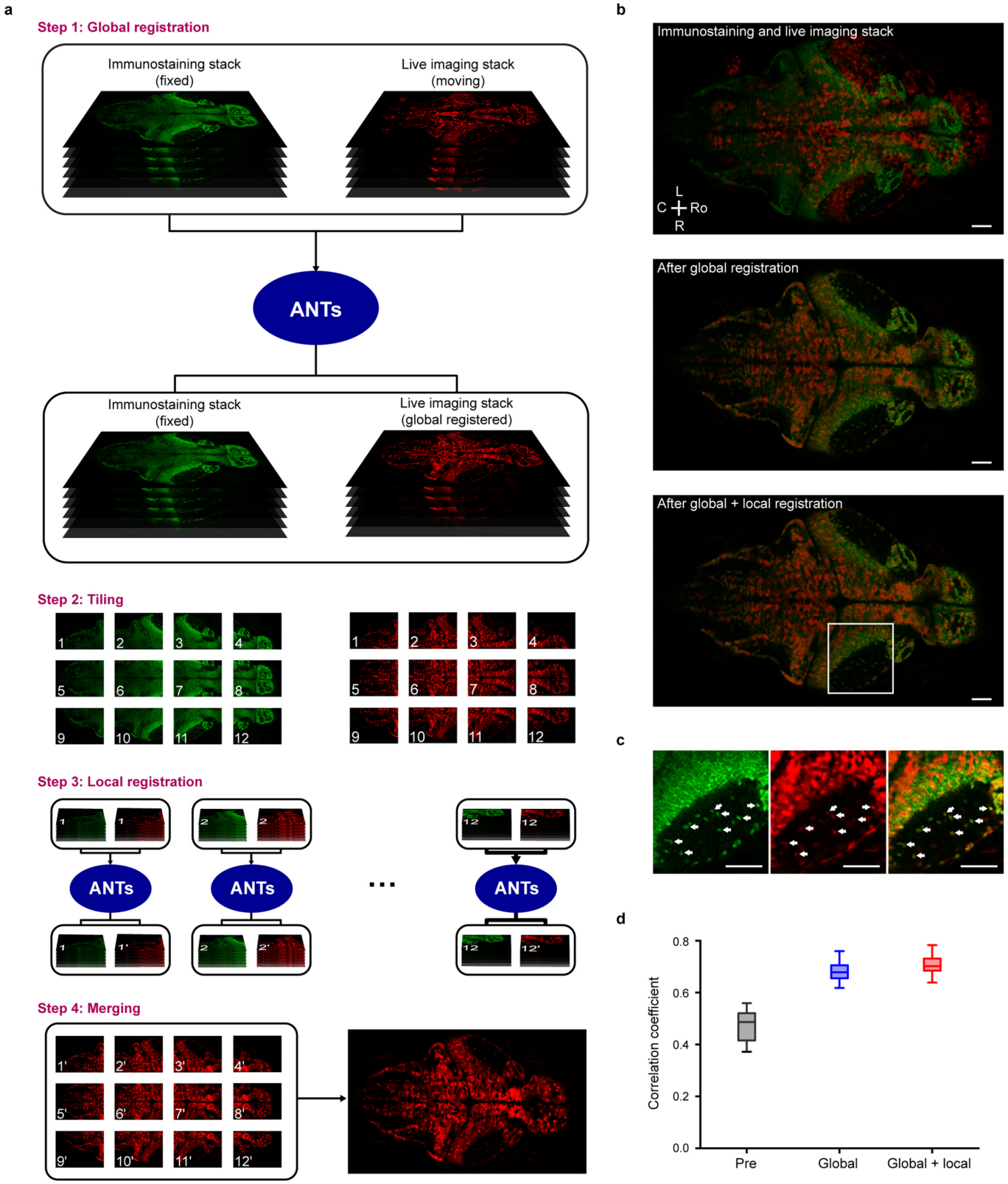
Extended Data Fig. 6 | PCA analysis of whole-brain neural activity across four fish. Each row shows a different animal. The first column shows all neuronal centroids identified from NMF (Methods) for each animal. The second column shows PCA trajectories in the first three PCA dimensions. The third column shows the locations of the exploitation-state-encoding neurons for each animal. The fourth column shows whole-brain PCA trajectories colour-coded by the activity of exploitation-state-encoding neurons in each animal.



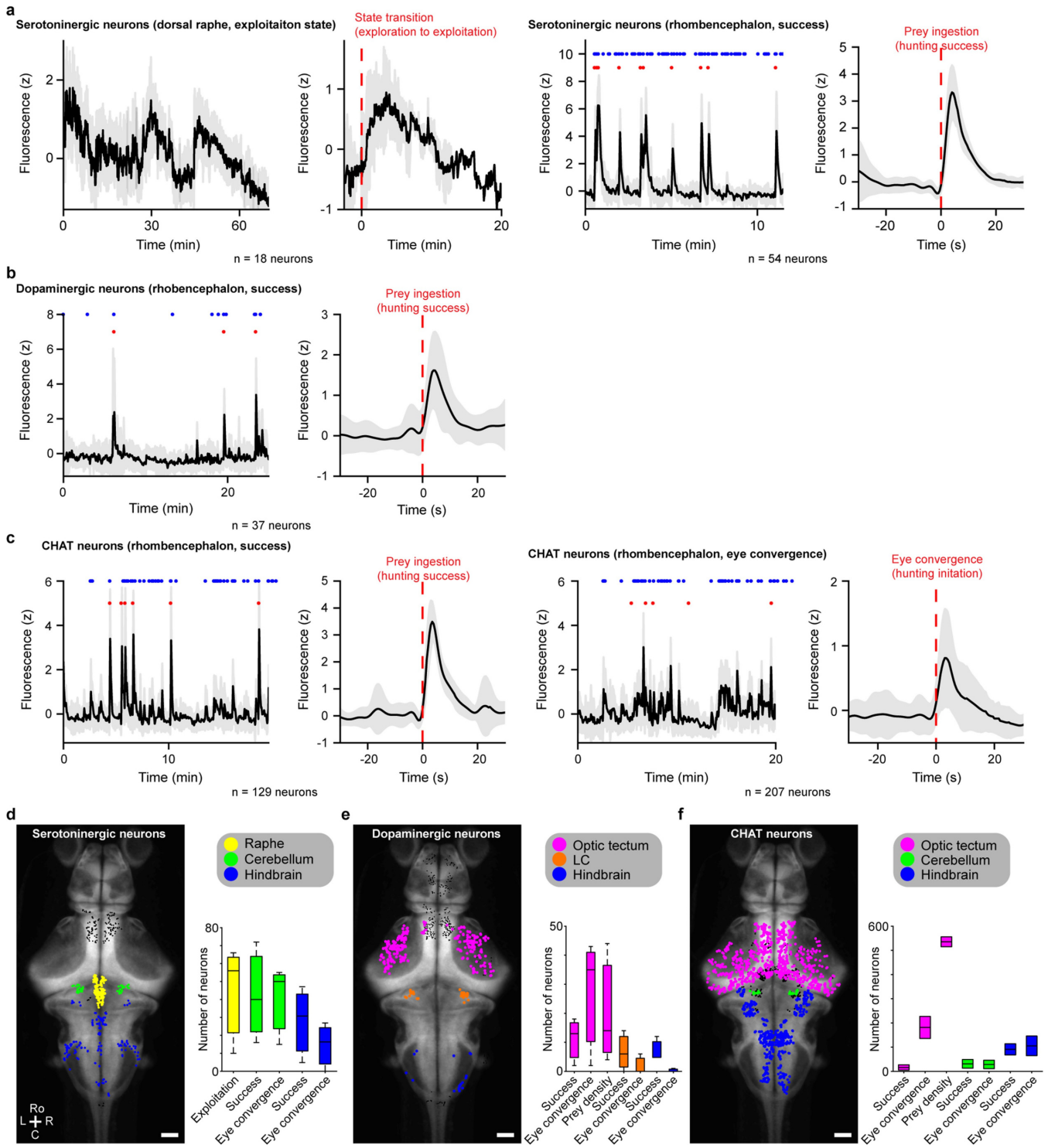
Extended Data Fig. 7 | Correlation between state duration and activity of dorsal raphe state neurons and exploration-state-encoding neurons, multimodal activation of the exploitation state, and stochastic model of trigger and state signals. a, Top, peak activities of exploration state neurons are not correlated with the duration of exploration state. Bottom, peak activities of dorsal raphe exploitation state encoding neurons correlate with the duration of exploitation state. Activity is in units of *z*-scored fluorescence. **b**–**d**, The trigger network that is active at the transition from exploration to exploitation contains both modality-specific neurons (**b**, **c**), that are only active at either spontaneous transitions (**b**) or at stimulus-evoked transitions (**c**), as well as cross-modal neurons that are activated at all state transitions regardless of whether the transition is stimulus-evoked or spontaneous (**d**). **e**, Schematic of the stochastic model. Spontaneous and stimulus-evoked trigger events are represented by delta functions with variable (integrated) amplitudes. The spontaneous trigger events arrive according to a Poisson process during exploration state, triggering the exit from exploration state after an exponentially distributed waiting time. Each delta function causes the trigger signal to undergo an impulsive rise, followed by rapid exponential decay. The state signal integrates the trigger signal, and then undergoes a slow linear decay back to baseline. Exploitation state corresponds to the time intervals when the state signal is elevated above baseline. Spontaneous trigger amplitudes are exponentially distributed, which leads to a corresponding amplitude in the state signal, and effectively determines the duration of the ensuing exploitation state due to the slow linear decay of the state signal. **f**, Two examples of recorded trigger and state signals during long exploitation states, together with corresponding signals in the fitted model. Event amplitudes A_i^{spont} and times t_i^{spont} are fit individually for each transition. The shared model parameters (α, β, γ) were obtained by simultaneous least squares fitting of the equations for $T(t)$ and $S(t)$ to the dynamics of the population mean activity of trigger and exploitation state neurons during 14 exploitation states having a duration of at least 10 min (obtained from $n = 17$ fish). The joint fitting procedure yielded $\alpha = 1.5$, $\beta = 0.1\,\text{min}^{-1}$, and $\gamma = 1.3\,\text{min}^{-1}$. These parameters, together with the fitted behavioural state distributions from Fig. 1d (obtained from $n = 36$ fish), comprise the conceptual model of the stochastic nonlinear dynamical system for exploration and exploitation.



Extended Data Fig. 8 | Validation of exploitation-state-encoding dorsal raphe neurons. a, **b**, Exploitation-state-encoding neurons have stable correlation to state even after excluding activity around hunting events. **a**, To verify the hypothesis that spatially significant (spatial *P* value < 0.025) exploitation-state-encoding neurons encode a persistent state and not simply hunting events, we removed the neural activity around all hunting events across the entire experiment, with increasing time intervals ranging from 0.5 s to 10 s in steps of 0.5 s. **b**, State-encoding neurons remain stably correlated to state after removal of increasingly large time windows around hunting events. Orange, spatially significant (spatial *P* value < 0.025) dorsal raphe exploitation-state-encoding neurons after spatial *P* value significance test. Blue, all spatially significant (spatial *P* value < 0.025) exploitation-state-encoding neurons ($n = 17$ animals). Standard error of all exploitation-state-encoding neurons is < 0.0024, and standard error of dorsal raphe exploitation-state-encoding neurons is < 0.0041. **c**–**f**, Validation of exploitation-state-encoding dorsal raphe neurons by clustering analysis. As a control test, we also conducted the same analysis on another brain region, the inferior olive (**g**–**j**). **c**, **g**, Clustering analysis of dorsal raphe neurons (**c**) and inferior olive neurons (control) (**g**). **d**, **h**, Activity of the tightest-correlated cluster (lowest average intra-cluster distance) in the dorsal raphe (**d**) and inferior olive (**h**) of each animal aligned to state transition from exploration to exploitation. **e**, **i**, Cross-validation of exploitation-state-encoding activity across animals for dorsal raphe (**e**) and inferior olive neurons (**i**). Each element of the array represents the mean correlation coefficient C_{ij} between the neural activity of the tightest-correlated dorsal raphe cluster of fish *i* and behaviour state of fish *j*. **f**, Statistical analysis between dorsal raphe neurons in on-diagonal elements ($i = j$) and neurons in off-diagonal elements ($i \neq j$). The mean correlation coefficient C_{ij}^k for $i = j$ was 0.468, whereas the mean correlation coefficient C_{ij}^k for $i \neq j$ was 0.024. A two-sample *t*-test analysis verified that there was a significant difference between these two groups ($P < 10^{-8}$, $n = 17$ animals). **j**, The same statistical analysis showed there was no significant difference between neurons in on-diagonal elements and neurons in off-diagonal elements for the inferior olive (mean correlation coefficient C_{ij}^k for $i = j$ was -0.047, mean correlation coefficient C_{ij}^k for $i \neq j$ was 0.023 ($P = 1$, two-sample *t*-test).



Extended Data Fig. 9 | Registration of live fluorescent brain volume to immunostained brain volume using ANTs. a, Schematic of ANTs registration pipeline (Methods). **b**, The superimposed live fluorescent and immunostained brain volumes before registration, after global registration, and after both global and local registration. **c**, The immunostained brain volume (green), live fluorescent brain volume (red), and their superimposed volume for a region in the optic tectum. The white arrows highlight the precise overlap of a sparse subpopulation of neurons in the optic tectum. **d**, Correlation between live fluorescent and immunostained brain volumes before and after registration (Methods). Scale bars, 50 µm (**b**, **c**).



Extended Data Fig. 10 | Activity of neuromodulatory subpopulations as a function of internal state transitions and hunting events. a–**c**, For each neuromodulatory population, we show an example trace of activity (left) and event-triggered average activity (right) aligned to state transition, successful ingestion of prey and/or hunting initiation. Blue dots, eye convergence or hunting initiation; red dots, prey ingestion. **d**–**f**, Left, spatial distribution of each neuromodulatory population. Neurons in five brain regions (optic tectum, raphe, locus coeruleus, cerebellum and hindbrain) are highlighted in different colours. Right, box plots showing, for each neuromodulatory population and brain region, the number of neurons that belong to each functional type. Scale bars, 50 µm (**d**–**f**). All plots report mean ± s.e.m.

nature research

Corresponding author(s): Drew N Robson & Jennifer M Li

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	Software for data collection was written in Julia 0.6.0 and C++11. GPU processing was implemented for fish tracking, as well as online and offline image registration. All code are already available from our previous publication.
Data analysis	All software for data analysis was written in Julia 0.6.0 and Python using the package ecosystem for both languages. All software for behavioral and image analysis, as well as license information and implementation notes, will be available online upon publication.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The data that support the findings of this study are available from the corresponding authors upon request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For behavioral analysis, a sample size of 7-36 animals was used depending on the experimental condition or analysis method. For neural data, a total of 17 animals was used.
Data exclusions	N/A
Replication	N/A
Randomization	N/A
Blinding	N/A

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Primary antibodies used to label dopamine neurons were mouse anti-tyrosine hydroxylase (Immunostar 22941). Primary antibodies used to label serotonin neurons were rabbit anti-5HT (serotonin) (Immunostar 20080). Primary antibodies used to label acetylcholine neurons were goat anti-Choline Acetyltransferase (Millipore AB144P).
Secondary antibodies used were donkey anti-mouse Alexa647 (Jackson ImmunoResearch 715-606-151), donkey anti-rabbit Alexa594 (Jackson ImmunoResearch 711-586-152) and donkey anti-goat Alexa647 (Jackson ImmunoResearch 705-606-147).

Validation

We validated immunostaining patterns for dopamine (anti-tyrosine hydroxylase), serotonin (anti-5HT), and CHAT (anti-Choline Acetyltransferase) against previous publications:

Dopamine:
Ryu S, Holzschuh J, Mahler J, Driever W. Genetic analysis of dopaminergic system development in zebrafish. J Neural Transm Suppl. 2006; 70:61–66.
Filippi A, Mahler J, Schweitzer J, Driever W. Expression of the paralogous tyrosine hydroxylase encoding genes th1 and th2 reveals the full complement of dopaminergic and noradrenergic neurons in zebrafish larval and juvenile brain. J Comp Neurol. 2010; 518:423–438. [PubMed: 20017209

Serotonin:
Norton WH, Folchert A, Bally-Cuif L. Comparative analysis of serotonin receptor (HTR1A/HTR1B families) and transporter (slc6a4a/b) gene expression in the zebrafish brain. J Comp Neurol. 2008; 511:521–542. [PubMed: 18839395]

Choline acetyltransferase (CHAT):
Hong E, Santhakumar K, Akitake CA, Ahn SJ, Thisse C, Thisse B, Wyart C, Mangin JM, Halpern ME. Cholinergic left-right asymmetry in the habenulo-interpeduncular pathway. Proc Natl Acad Sci USA. 2013; 110:21171–21176. [PubMed: 24327734]

As a collective group, to get a better idea of the 3 dimensional expression patterns for all three neuromodulators, we further referenced:
P. Panula, Y.C. Chen, M. Priyadarshini, H. Kudo, S. Semenova, M. Sundvik, V. Sallinen. The comparative neuroanatomy and neurochemistry of zebrafish CNS systems of relevance to human neuropsychiatric diseases. Neurobiology of Disease 2010

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	zebrafish larvae (6-8 dpf). Sexual differentiation has not occurred by this age.
Wild animals	N/A
Field-collected samples	N/A

A bacteriophage nucleus-like compartment shields DNA from CRISPR nucleases

https://doi.org/10.1038/s41586-019-1786-y

Received: 16 July 2018

Accepted: 11 October 2019

Published online: 9 December 2019

Senén D. Mendoza[1], Eliza S. Nieweglowska[2,7], Sutharsan Govindarajan[1,5,7], Lina M. Leon[1], Joel D. Berry[1], Anika Tiwari[1], Vorrapon Chaikeeratisak[3,6], Joe Pogliano[3], David A. Agard[2,4] & Joseph Bondy-Denomy[1,4]*

All viruses require strategies to inhibit or evade the immune pathways of cells that they infect. The viruses that infect bacteria, bacteriophages (phages), must avoid immune pathways that target nucleic acids, such as CRISPR–Cas and restriction-modification systems, to replicate efficiently[1]. Here we show that jumbo phage ΦKZ segregates its DNA from immunity nucleases of its host, *Pseudomonas aeruginosa*, by constructing a proteinaceous nucleus-like compartment. ΦKZ is resistant to many immunity mechanisms that target DNA in vivo, including two subtypes of CRISPR–Cas3, Cas9, Cas12a and the restriction enzymes HsdRMS and EcoRI. Cas proteins and restriction enzymes are unable to access the phage DNA throughout the infection, but engineering the relocalization of EcoRI inside the compartment enables targeting of the phage and protection of host cells. Moreover, ΦKZ is sensitive to Cas13a—a CRISPR–Cas enzyme that targets RNA—probably owing to phage mRNA localizing to the cytoplasm. Collectively, we propose that *Pseudomonas* jumbo phages evade a broad spectrum of DNA-targeting nucleases through the assembly of a protein barrier around their genome.

Phages that infect *Pseudomonas aeruginosa* can avoid destruction mediated by CRISPR by encoding 'anti-CRISPR' proteins that inhibit the type I-E and I-F CRISPR–Cas systems[2–4]. To determine whether any *P. aeruginosa* phages are resistant to the *P. aeruginosa* type I-C CRISPR–Cas system[5] (a common, but understudied variant[6]), we engineered a strain of *P. aeruginosa* to express type I-C *cas3*, *cas5*, *cas8* and *cas7*[7], and provided this strain with a panel of CRISPR RNAs (crRNAs) that target phages from five taxonomic groups: JBD30, D3, JBD68 (distinct temperate siphophages), F8 and ΦKZ (distinct lytic myophages). All phages succumbed to targeting (Fig. 1a), except for ΦKZ (Fig. 1b) and a related phage, ΦPA3 (Extended Data Fig. 1a). The ΦKZ titre did not decrease when ΦKZ was challenged with 11 different type I-C crRNAs (Fig. 1b, Extended Data Fig. 2a), nor when it was targeted by the type I-F CRISPR–Cas system of *P. aeruginosa*[8] (Extended Data Fig. 2b).

Phage ΦKZ resists CRISPR–Cas targeting

The ΦKZ genome has no homologues of known anti-CRISPR (Acr) genes[2–4,9,10], or anti-CRISPR-associated genes, which have previously enabled identification of Acr genes[4,9,10]. Moreover, gene knockout approaches have not been established for ΦKZ. Therefore, to determine the mechanism by which this phage can evade CRISPR-mediated destruction we used type II-A CRISPR–Cas9 from *Streptococcus pyogenes* (*Spy*Cas9). *Spy*Cas9 and single-guide RNAs (sgRNAs) robustly targeted control phage JBD30, but phage replication and the associated lysis of host cells were unaffected during phage ΦKZ infection (Fig. 1c). An additional six sgRNA sequences were also unable to impact ΦKZ replication (Extended Data Fig. 3a), as were four sgRNAs against ΦPA3 (Extended Data Fig. 1b). Given the ability of this phage to evade unrelated CRISPR systems (types I and II)—including one from a microorganism that this phage does not infect (*S. pyogenes*)—we hypothesized that ΦKZ might be generally resistant to CRISPR–Cas immunity, as opposed to relying on specific inhibitor proteins. When we expressed the type V-A Cas12a (Cpf1) CRISPR–Cas system from *Moraxella bovoculi* in *P. aeruginosa*, the control phage was targeted successfully, but ΦKZ was not, with any of the nine crRNAs tested (Fig. 1d, Extended Data Fig. 3b). The ability of ΦKZ to resist both CRISPR systems that are found in its natural host, *Pseudomonas* (types I-C and I-F), and those that are not naturally present in *Pseudomonas* (types II-A and V-A) suggests that this phage has a mechanism that enables 'pan-CRISPR' resistance.

Restriction-modification (R-M) systems are the most common form of bacterial immunity in nature and pose a substantial impediment to phage replication[1]. We next tested type I and type II R-M systems (HsdRMS from *P. aeruginosa* and EcoRI from *Escherichia* coli, respectively; 24 and 92 cut sites in the ΦKZ genome, respectively). ΦKZ was propagated on the *P. aeruginosa* strain PAK, an isolate that generates phages that are susceptible to HsdRMS restriction when infecting the PAO1 strain. When phage JBD30 (which has five type I R-M sites) was assayed in this manner, its titre was reduced by five orders of magnitude, which was dependent on *hsdR* (Fig. 1e). By contrast, no restriction was observed for ΦKZ (Fig. 1e) or ΦPA3 (Extended Data Fig. 1c). Similarly, the expression of EcoRI reduced the titre of JBD30 (6 EcoRI sites) by three orders of magnitude, but had no effect on ΦKZ (92 EcoRI sites) (Fig. 1f). Together, these experiments demonstrate that ΦKZ is

[1]Department of Microbiology and Immunology, University of California San Francisco, San Francisco, CA, USA. [2]Howard Hughes Medical Institute, Department of Biochemistry and Biophysics, University of California San Francisco, San Francisco, CA, USA. [3]Division of Biological Sciences, University of California San Diego, La Jolla, CA, USA. [4]Quantitative Biosciences Institute, University of California San Francisco, San Francisco, CA, USA. [5]Present address: Department of Biology, SRM University AP, Amaravati, India. [6]Present address: Department of Biochemistry, Faculty of Science, Chulalongkorn University, Bangkok, Thailand. [7]These authors contributed equally: Eliza S. Nieweglowska, Sutharsan Govindarajan. *e-mail: Joseph.Bondy-Denomy@ucsf.edu

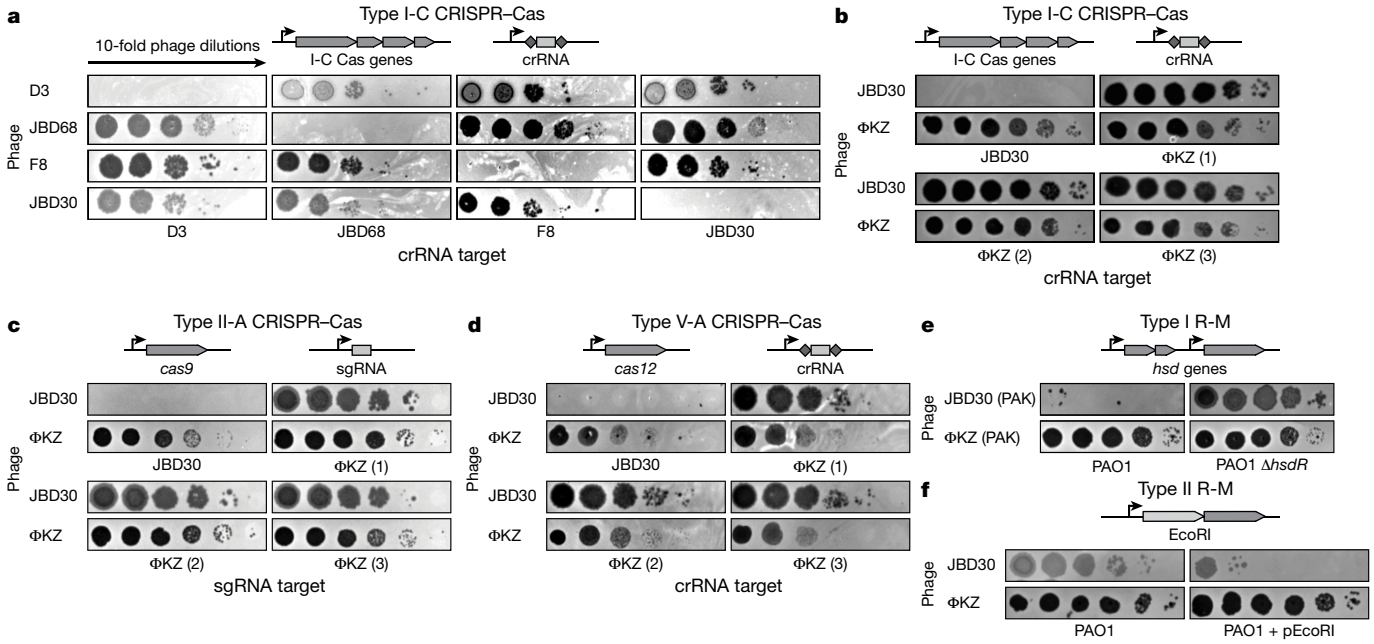


Fig. 1 | Identification of a phage that resists targeting by diverse CRISPR–Cas and R-M systems. Plaque assays with the indicated phage spotted in tenfold serial dilutions on a lawn of *P. aeruginosa*; dark clearings in the lawn represent phage replication. **a**–**d**, Strain PAO1 expressing type I-C Cas genes (*cas3*, *cas5*, *cas8*, *cas7*) and crRNAs that target the indicated phages (**a**); type I-C Cas genes and crRNAs that target phage JBD30 or distinct crRNAs (1–3) that target phage ΦKZ (**b**); type II-A *cas9* and distinct sgRNAs that target the indicated phages (**c**); or type V-A *cas12a* and distinct crRNAs against the indicated phages (**d**). **e**, An endogenous type I R-M system (hsdRSM) in wild-type PAO1 and a mutant strain with an isogenic *hsdR* knockout (PAO1Δ*hsdR*) was assayed using phages that were propagated on PAK (for example, JBD30 (PAK) was first propagated on strain PAK). **f**, The type II EcoRI R-M system in PAO1. Restriction activity was assayed using phages JBD30 and ΦKZ. pEcoRI denotes plasmid-expressed EcoRI. All plaque assays were replicated two or more times with similar results.

refractory to our six selected CRISPR–Cas (types I-C, I-F, II-A, and V-A) and restriction endonucleases (types I and II) in vivo.

Protein barrier occludes immune enzymes

ΦKZ and ΦKZ-like phages that infect *P. aeruginosa* and *Pseudomonas chlororaphis* have been shown to construct an elaborate proteinaceous nucleus-like compartment in which the replication of phage DNA takes place[11,12]. In addition, a phage-encoded homologue of tubulin, PhuZ, centres the compartment within the host cell[11–15]. Proteins that are involved in DNA replication, transcription and recombination localize inside the shell, whereas mRNA and proteins that mediate translation localize in the cytoplasmic space[11]—similarly to the eukaryotic nucleus. Given the apparent exclusion of specific proteins, we hypothesized that this structure might be responsible for the pan-resistance of ΦKZ to distinct immune processes.

P. aeruginosa cells that were infected with ΦKZ were imaged with immunofluorescence to detect Cas9 (Fig. 2a, Extended Data Fig. 4). We also used live-cell imaging to detect Cherry fusions with Cas9, two signature proteins from the type I-C and I-F CRISPR–Cas systems (Cas8 and Cas3; Fig. 2b) and the restriction enzyme HsdR (Fig. 2c). These experiments showed that the immune enzymes are excluded from the shell during phage infection. DAPI staining revealed that the phage DNA was located inside the shell, whereas the host genome was degraded[14]. The phage protein ORF152 (imaged with anti-Myc immunofluorescence and Cherry fusion) and host protein topoisomerase I (Cherry fusion)—both of which have been previously shown to be internalized in the shell—colocalized with the DAPI signal, but Cherry was excluded (Fig. 2d). Although the rules that govern protein internalization in the shell are currently unknown, each protein of known function that localizes inside the shell interacts with DNA[11,12] (that is, forms part of the DNA replication and transcription machinery), suggesting that the

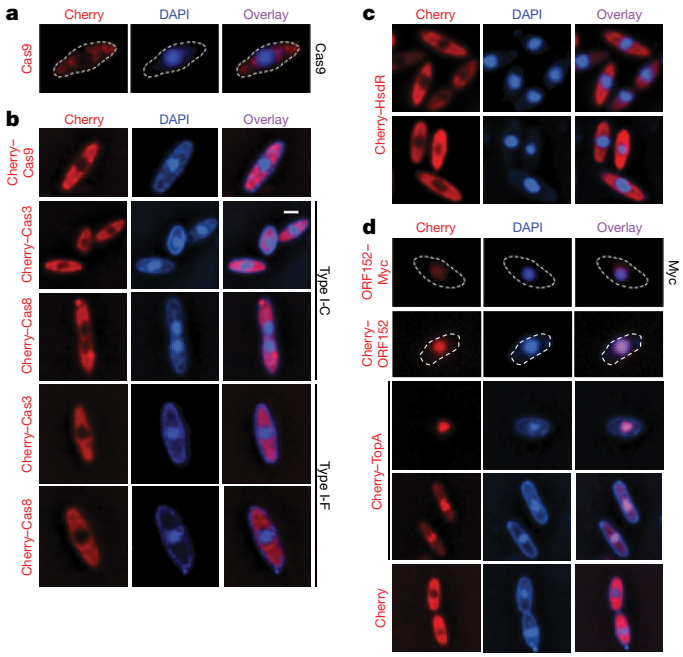


Fig. 2 | CRISPR–Cas proteins and restriction enzymes are excluded from the nucleus-like structure of ΦKZ. a, Fluorescence microscopy of *P. aeruginosa*, immunostained for Cas9. The DAPI stain shows the phage DNA within the nucleus-like structure. **b**, **c**, Live-cell fluorescence imaging of *P. aeruginosa* strains that were engineered to express type II-A Cas9, type I-C or I-F Cas8 or Cas3 proteins fused to Cherry (**b**) or HsdR fused to Cherry (**c**). Scale bar, 1 μm. **d**, Top row, fluorescence microscopy of Myc-tagged ORF152 in *P. aeruginosa*, imaged with anti-Myc immunofluorescence. Bottom rows, live-cell fluorescence imaging of ORF152 and TopA proteins fused to Cherry, or Cherry alone. All experiments were replicated two or more times with similar results.

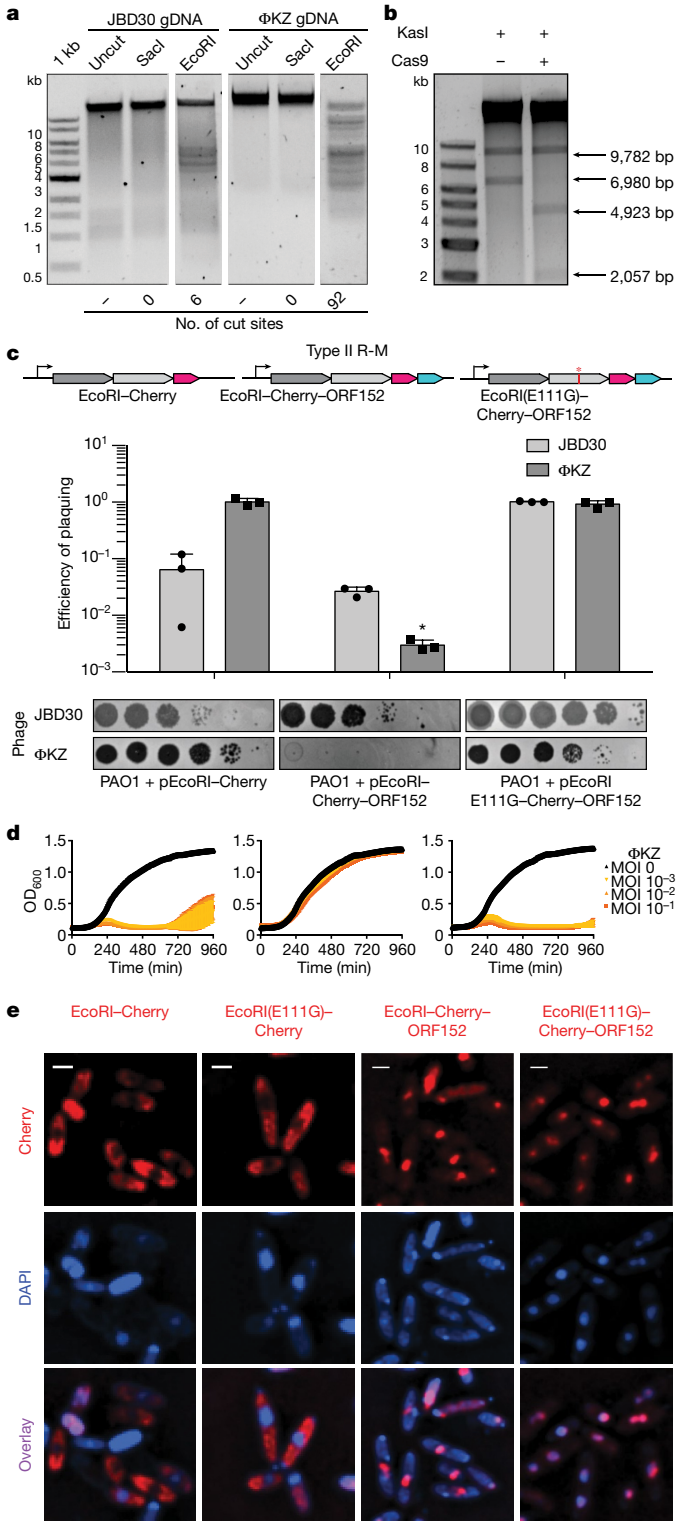


Fig. 3 | ΦKZ gDNA can be cleaved by immune enzymes. a, ΦKZ and JBD30 gDNA digested with the indicated restriction enzymes in vitro. The first lane contains a 1-kb DNA ladder. The number of cut sites for each enzyme is shown at the bottom of the gels. Products were visualized on a 0.7% agarose gel stained with the DNA stain SYBR Safe. **b**, ΦKZ gDNA digested in vitro using KasI, and incubated with or without Cas9 loaded with crRNA:tracrRNA that targets the fragment liberated by KasI. Products were visualized on a 0.7% agarose gel stained with SYBR Safe. **c**, Plaque assays for strain PAO1 expressing the fusion proteins EcoRI–Cherry, EcoRI–Cherry–ORF152 or EcoRI(E111G)–Cherry–ORF152. Plaque assays were conducted as in Fig. 1a and quantified ($n = 3$). Data are mean ± s.d. A Student's *t*-test comparing the efficiency of plaquing of ΦKZ on PAO1 that expresses pEcoRI–Cherry to PAO1 that expresses pEcoRI–Cherry–ORF152 yielded a *P* value of 2.88×10^{-4} (indicated with an asterisk). **d**, Growth curves measuring the optical density at 600 nm ($OD_{600\,nm}$) of PAO1 cells that were infected with ΦKZ at the indicated multiplicity of infection (MOI). **e**, Live-cell fluorescence imaging of infected strains of *P. aeruginosa* that were engineered to express EcoRI–Cherry, EcoRI(E111G)–Cherry, EcoRI–Cherry–ORF152 or EcoRI(E111G)–Cherry–ORF152. The DAPI stain shows the phage DNA. The in vitro digestion experiments in **a**, **b** were replicated twice with similar results. The plaque assays (**c**), growth curves (**d**) and microscopy (**e**) were replicated three times. Scale bars, 1 µm.

exclusion of DNA-binding Cas and restriction proteins is an adaptive function of the shell.

To confirm that the ΦKZ phage genome could be a substrate for DNA cleavage, if accessed, two enzymes that did not cleave ΦKZ in vivo, EcoRI and Cas9, were assayed in vitro. ΦKZ DNA was extracted from virions and subjected to restriction digestion reactions with a panel of restriction enzymes. EcoRI, HindIII, KpnI and NcoI all cleaved the DNA, whereas SacI—which lacks a sequence-recognition motif in the ΦKZ genome—did not (Fig. 3a, Extended Data Fig. 5a). The ability of Cas9 to cleave ΦKZ genomic DNA (gDNA) in vitro was next assessed. Owing to the large size of the ΦKZ genome (280 kb), we first subjected purified phage DNA to the restriction enzyme KasI to yield a 6.98-kb product, and then cleaved that species with a *Spy*Cas9 nuclease loaded with both crRNA and trans-activating crRNA (tracrRNA) in vitro. This reaction depleted the substrate and produced the expected 4.9-kb and 2.1-kb fragments, confirming that the phage gDNA is sensitive to Cas9 cleavage (Fig. 3b). Together, these results demonstrate that immune enzymes are capable of cleaving ΦKZ gDNA when they can access it, and that immune evasion is probably not a result of an intrinsic feature of the phage DNA (such as base modifications that can impede the Cascade–Cas3 complex, Cas9 and EcoRI)[16–19].

Phage targeting via enzyme localization

As there is no currently known method that can be used to mutate, weaken or knockout the phage shell structure, we next sought to enable an immune enzyme to bypass the shell and thereby access the phage DNA in vivo. We fused the single effector enzyme Cas9 to ORF152, a phage-encoded RecA-like protein that is internalized within the shell[11,12]. Independent fusions of Cas9 to the N or the C terminus of ORF152 did not affect the replication of ΦKZ (Extended Data Fig. 6a, b). Imaging of one orientation of ORF152–Cas9 revealed peri-shell localization (Extended Data Fig. 6c, d); this suggests that the fusion redirected Cas9 from its previously diffuse state but that the large Cas9 protein (1,368 amino acids, 158 kDa) was unable to traverse the shell border.

Given the large size (for example, Cas12a) and complexity (for example, Cascade, HsdRMS) of each of the other immune effectors, we next fused the small, single effector protein EcoRI (278 amino acids, 31.5 kDa) to Cherry-tagged ORF152. This fusion resulted in a notable decrease of more than 2.5 orders of magnitude in the ΦKZ titre and markedly reduced plaque sizes (Fig. 3c. Liquid infections revealed that cells that express EcoRI–Cherry–ORF152 were extremely well protected, gaining an increase of more than five orders of magnitude in their resistance to phage-induced lysis (Fig. 3d). A catalytic mutant EcoRI(E111G) fused to Cherry–ORF152 exhibited no immune activity against ΦKZ in either assay, and nor did active EcoRI–Cherry without the ORF152 fusion (Fig. 3c, d). Similar results were observed in the absence of the Cherry tag (that is, EcoRI–ORF152); however, ΦKZ targeting was more modest in that case, reducing titres by approximately tenfold and protecting cells by a factor of around 10^4 (Extended Data Fig. 7). Imaging of infected cells that express EcoRI–Cherry or EcoRI(E111G)–Cherry demonstrated that these proteins are excluded from the shell, whereas EcoRI(E111G)–Cherry–ORF152 is successfully localized inside (Fig. 3e;

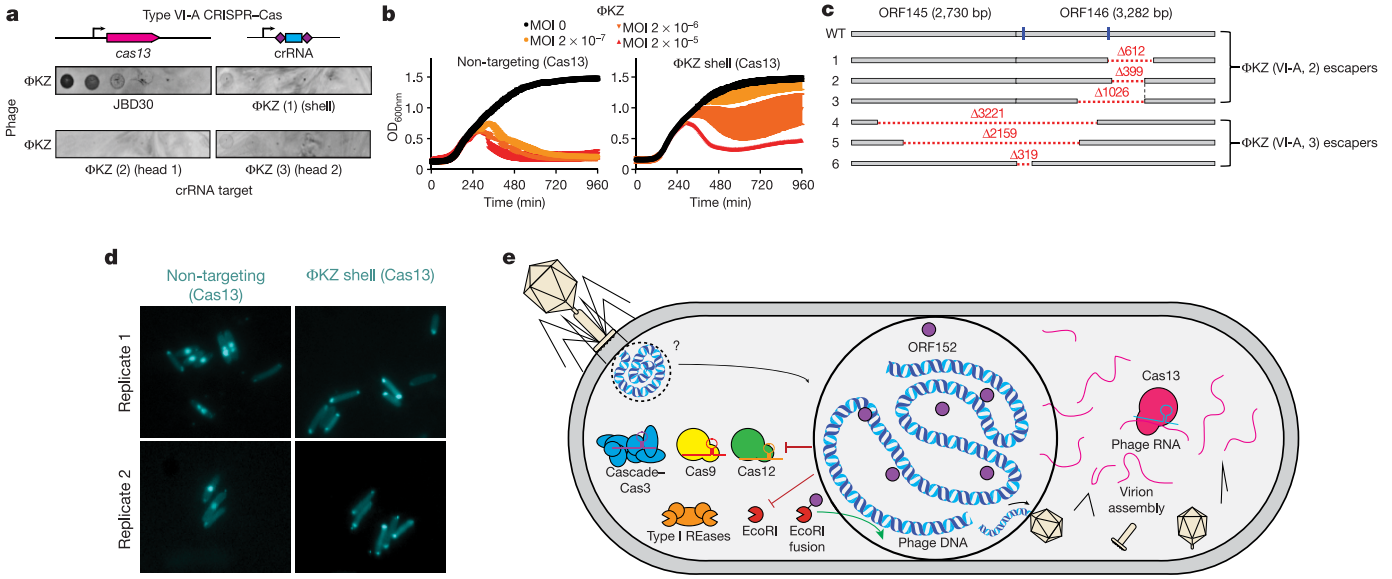


Fig. 4 | Phage ΦKZ DNA is sensitive to RNA targeting with Cas13. a, Strain PAO1 expressing *Lse*Cas13a and distinct crRNAs that target the shell (one crRNA) or the head (two crRNAs) of phage ΦKZ. Plaque assays were conducted as in Fig. 1a. **b**, Growth curves measuring the $OD_{600\,nm}$ of PAO1 cells that were infected with ΦKZ at the indicated MOI. **c**, Wild-type (WT) and Cas13 escaper ΦKZ phage, with deletions indicated with red dashed lines and target sites with blue lines. **d**, Live-cell fluorescence imaging of *P. aeruginosa* strains that were engineered to express *Lse*Cas13a and crRNAs that target ΦKZ. The cyan stain shows the DAPI-stained phage DNA. **e**, Summary model. The ΦKZ nucleus-like structure excludes Cas9, Cas12, Cascade–Cas3 (type I-C and type I-F), type I restriction endonucleases (REase) and type II restriction endonucleases (for example, EcoRI); however, the mRNA (red) is exported and can be targeted by Cas13. The nucleus-like structure is resistant to the indicated nucleases, but fusing EcoRI to the internal protein ORF152 enables targeting. The assays in **a** were replicated more than three times with similar results. The growth curve experiments (**b**) were replicated twice with similar results. Escaper mutants (**c**) were isolated once and verified by PCR, sequencing and plaque assays. Microscopy (**d**) was replicated twice with similar results.

see Supplementary Video 1 for time-lapse images). EcoRI–Cherry–ORF152 impaired the ability of infected cells to form full shells and proceed through the infection process, and the EcoRI fusion protein was often localized within or adjacent to the DAPI-stained puncta (Fig. 3e, Supplementary Video 2). Some host DNA can be seen in these cells and infection does not proceed (Fig. 3e, Extended Data Fig. 7c, d). By bypassing the physical barrier of the shell using rational engineering, we conclude that the shell is the cause of the resistance to immunity that is shown by jumbo phage ΦKZ.

Phage mRNA is sensitive to Cas13a

The nucleus-like structure produced by ΦKZ provides robust resistance to DNA-targeting immune pathways, but other immune systems may exist in vivo that can evade this mechanism. We envisaged that the mRNA that is exported for translation in the cytoplasm is probably susceptible to targeting and that a ribonuclease could provide anti-ΦKZ immunity. To test this, we adapted the type VI-A CRISPR RNA-guided RNA nuclease Cas13a from *Listeria seeligeri* (*Lse*Cas13a) [20,21] for phage targeting in *P. aeruginosa*. Three *Lse*Cas13a spacers (two targeting the head gene *gp146* and one targeting the shell gene *gp054*) decreased the efficiency of ΦKZ plaquing by more than 10^6 fold (Fig. 4a, Extended Data Fig. 8). Corroborating these results, *Lse*Cas13a was also effective at protecting *P. aeruginosa* in liquid cultures, with an increase of 1–2 orders of magnitude in resistance to phage-induced lysis (Fig. 4b). We suspect that the previously demonstrated collateral damage caused by Cas13a[22] is detrimental to cell fitness, which leads to an enhancement of the growth curves that is modest compared to that brought about by EcoRI targeting. Cas13 escaper phages (that is, phages that evaded Cas13-mediated targeting) were also identified (Extended Data Fig. 9); these occurred at a frequency of around 10^{-6} to 10^{-7} and, surprisingly, contained genomic deletions at the site of targeting (Fig. 4c). Notably, all deletions were in-frame and five out of six escaper phages had clear regions of microhomology flanking the deletion that ranged from 7–21 bp in length. One guide RNA targeting the 5′ end of *orf146* selected for deletions in which the upstream gene (*orf145*) became fused, in frame, to the targeted gene (Fig. 4c). These data support the conclusion that sequence-specific RNA targeting by CRISPR–Cas proteins can inhibit this phage.

Before the construction of the shell takes place, the state of the phage DNA is unknown. During Cas13a targeting of the shell mRNA, imaging revealed that infections arrested before the phage DNA proceeded from its injection site at the poles. Most cells had two polar puncta, even one hour after infection (Fig. 4d). The absence of diffusion or clearance of the phage DNA (for example, by the endogenous type I R-M system) suggests that the injected phage genome may be protected before shell assembly by injected phage proteins or pre-existing host factors. The potential for a DNA structure to exist prior to shell construction is further corroborated by our observation that in the absence of Cas13a-mediated arrest, DAPI-stained phage DNA was detected adjacent to EcoRI–Cherry–ORF152 puncta (Fig. 3e) or adjacent to a nascent shell in which TopA was localized (Extended Data Fig. 10). The molecular nature and mechanisms underlying this early protection event remain to be elucidated.

Discussion

The assembly of a proteinaceous compartment in which phage DNA replication occurs creates a physical protective barrier that results in the resistance of phage ΦKZ to enzymes that cleave DNA (Fig. 4e). The evasion of endogenous *P. aeruginosa* type I CRISPR–Cas systems by ΦKZ suggests that these jumbo phages are probably resistant to type I CRISPR–Cas systems in nature. Supporting this hypothesis, our analysis of a previously published collection of over 4,000 non-redundant *P. aeruginosa* spacers[5] (types I-C, I-E and I-F) found no spacers against ΦKZ or its jumbo phage relatives ΦPA3, PaBG, KTN4 and PA7

Table 1 | No spacers against ΦKZ-like phages

Phage	No. of spacers	Type I CRISPR sensitivity	Reference
DMS3m	75	I-C: sensitive; I-E: resistant (AcrIE3); I-F: sensitive	This study and previous studies[3,8]
JBD30	51	I-C: sensitive; I-E: sensitive; I-F: resistant (AcrIF1)	This study and previous studies[2,3]
JBD18	51	I-F: sensitive	A previous study[8]
JBD25	46	I-F: sensitive	A previous study[8]
JBD68	28	I-C: sensitive	This study
D3	49	I-C: sensitive	This study
F8	3	I-C: sensitive	This study
ΦKZ	0	I-C: resistant; I-F: resistant	This study
phiPA3	0	I-C: resistant	This study
PaBG	0	not assayed	
KTN4	0	not assayed	
PA7	0	not assayed	

The total number of type I CRISPR spacers with a perfect match to the indicated *P. aeruginosa* phages assayed in this study and previous CRISPR–Cas studies. The experimental sensitivity of each phage to the indicated subtypes is shown. AcrIE3 and AcrIF1 are I-E and I-F anti-CRISPR proteins, respectively. For F8, the 3 spacers have mismatches (4 nucleotides or fewer) to the F8 genome.

(Table 1). By contrast, many spacers from each system match diverse *P. aeruginosa* phages, such as those assayed in our screen (Fig. 1) and those encoding anti-CRISPR proteins. In addition, given the efficacy of the RNA-targeting CRISPR–Cas13 system, we propose that type VI CRISPR systems are well suited to target the mRNA of DNA phages when the DNA is inaccessible (that is, owing to base modifications or physical segregation).

Although this phage compartment has only been documented among the jumbo phages of *Pseudomonas*[11,12], physical occlusion of phage gDNA through this and other mechanisms may constitute a route through which immune systems can be evaded by other phages. Indeed, 'mega phages' have recently been reported to encode homologues of phage tubulin, which in *Pseudomonas* are involved in centring the proteinaceous compartment during infection[23]. The pan-resistance of ΦKZ to enzymes that target DNA provides an explanation for the elaborate, nucleus-like structure. Furthermore, the polar localization of the injected phage DNA during mRNA targeting suggests a poorly understood early protective mechanism. Considering the pronounced resistance to overexpressed immune enzymes and the previously observed docking of capsids at the shell periphery[11,24], we propose that the phage DNA is never exposed to the cytoplasm (Fig. 4e). Other hypotheses to explain the importance of the shell—including protection from phage-derived nucleases that degrade the bacterial genome, and spatial organization of the large phage genome during replication and packaging—remain to be addressed. Regardless, we conclude that the phage-assembled nucleus-like structure provides a strong protective barrier to DNA-targeting immune pathways.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1786-y.

1. Koonin, E. V., Makarova, K. S. & Wolf, Y. I. Evolutionary genomics of defense systems in archaea and bacteria. *Annu. Rev. Microbiol.* **71**, 233–261 (2017).
2. Bondy-Denomy, J., Pawluk, A., Maxwell, K. L. & Davidson, A. R. Bacteriophage genes that inactivate the CRISPR/Cas bacterial immune system. *Nature* **493**, 429–432 (2013).
3. Pawluk, A., Bondy-Denomy, J., Cheung, V. H. W., Maxwell, K. L. & Davidson, A. R. A new group of phage anti-CRISPR genes inhibits the type I-E CRISPR–Cas system of *Pseudomonas aeruginosa*. *MBio* **5**, e00896-e14 (2014).
4. Pawluk, A. et al. Inactivation of CRISPR–Cas systems by anti-CRISPR proteins in diverse bacterial species. *Nat. Microbiol.* **1**, 16085 (2016).
5. van Belkum, A. et al. Phylogenetic distribution of CRISPR–Cas systems in antibiotic-resistant *Pseudomonas aeruginosa*. *MBio* **6**, e01796-e15 (2015).
6. Makarova, K. S. et al. An updated evolutionary classification of CRISPR–Cas systems. *Nat. Rev. Microbiol.* **13**, 722–736 (2015).
7. Marino, N. D. et al. Discovery of widespread type I and type V CRISPR–Cas inhibitors. *Science* **362**, 240–242 (2018).
8. Cady, K. C., Bondy-Denomy, J., Heussler, G. E., Davidson, A. R. & O'Toole, G. A. The CRISPR/Cas adaptive immune system of *Pseudomonas aeruginosa* mediates resistance to naturally occurring and engineered phages. **194**, 5728–5738 (2012).
9. Pawluk, A. et al. Naturally occurring off-switches for CRISPR–Cas9. *Cell* **167**, 1829–1838 (2016).
10. Rauch, B. J. et al. Inhibition of CRISPR–Cas9 with bacteriophage proteins. *Cell* **168**, 150–158 (2017).
11. Chaikeeratisak, V. et al. Assembly of a nucleus-like structure during viral replication in bacteria. *Science* **355**, 194–197 (2017).
12. Chaikeeratisak, V. et al. The phage nucleus and tubulin spindle are conserved among large *Pseudomonas* phages. *Cell Rep.* **20**, 1563–1571 (2017).
13. Kraemer, J. A. et al. A phage tubulin assembles dynamic filaments by an atypical mechanism to center viral DNA within the host cell. *Cell* **149**, 1488–1499 (2012).
14. Erb, M. L. et al. A bacteriophage tubulin harnesses dynamic instability to center DNA in infected cells. *eLife* **3**, e03197 (2014).
15. Zehr, E. A. et al. The structure and assembly mechanism of a novel three-stranded tubulin filament that centers phage DNA. *Structure* **22**, 539–548 (2014).
16. Bryson, A. L. et al. Covalent modification of bacteriophage T4 DNA inhibits CRISPR–Cas9. *mBio* **6**, e00648-15 (2015).
17. Strotskaya, A. et al. The action of *Escherichia coli* CRISPR–Cas system on lytic bacteriophages with different lifestyles and development strategies. *Nucleic Acids Res.* **45**, 1946–1957 (2017).
18. Vlot, M. et al. Bacteriophage DNA glucosylation impairs target DNA binding by type I and II but not by type V CRISPR–Cas effector complexes. *Nucleic Acids Res.* **46**, 873–885 (2018).
19. Huang, L. H., Farnet, C. M., Ehrlich, K. C. & Ehrlich, M. Digestion of highly modified bacteriophage DNA by restriction endonucleases. *Nucleic Acids Res.* **10**, 1579–1591 (1982).
20. Abudayyeh, O. O. et al. C2c2 is a single-component programmable RNA-guided RNA-targeting CRISPR effector. *Science* **353**, aaf5573 (2016).
21. Gootenberg, J. S. et al. Nucleic acid detection with CRISPR-Cas13a/C2c2. *Science* **356**, 438–442 (2017).
22. Meeske, A. J. & Marraffini, L. A. RNA guide complementarity prevents self-targeting in type VI CRISPR systems. *Mol. Cell* **71**, 791–801 (2018).
23. Al-Shayeb, B. et al. Clades of huge phage from across Earth's ecosystems. Preprint at *bioRxiv* https://www.biorxiv.org/content/10.1101/572362v1 (2019).
24. Chaikeeratisak, V. et al. Viral capsid trafficking along treadmilling tubulin filaments in bacteria. *Cell* **177**, 1771–1780 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized, and investigators were not blinded to allocation during experiments and outcome assessment.

Bacterial growth and genetic manipulation

The strains, plasmids, phages and spacer sequences used in this study are listed in the Supplementary Information. *P. aeruginosa* strain PAO1 was grown in LB at 37 °C with aeration at 225 rpm. When necessary, plating was performed on LB agar with carbenicillin (250 μg ml^{-1}) or gentamicin (50 μg ml^{-1}). Gene expression was induced by the addition of L-arabinose (0.1% final) and/or isopropyl β-D-1-thiogalactopyranoside (IPTG; 0.5 mM or 1 mM final). For chromosomal insertions at the attTn7 locus, *P. aeruginosa* cells were electroporated with the integrating vector pUC18T-lac and the transposase expressing helper plasmid pTNS3, and selected on gentamicin. Potential integrants were screened by colony PCR with primers PTn7R and PglmS-down. Electrocompetent cell preparations, transformations, integrations, selections, plasmid curing and FLP-recombinase-mediated marker excision with pFLP were performed as described previously[25].

Phage growth and DNA extraction

Phage growth was conducted in LB at 37 °C with PAO1 as a host. Growth curves were conducted in a Biotek Synergy plate reader at 37 °C with orbital shaking set to maximum speed. Phage stocks were diluted, stored in SM buffer[8] and used for routine plaquing assays. Plaque assays were conducted at 37 °C with 20 ml of bottom agar containing 10 mM $MgSO_4$ and 0.35% or 0.7% top agar (often both concentrations were used in parallel) also containing 10 mM $MgSO_4$ and any inducer molecules. Spots of 3 μl were applied to the top agar after it had been poured and solidified. For high-titre lysates to generate phage DNA, plates with a high number of plaques were flooded with SM buffer and collected[8]. The lysates were subsequently treated with DNase. Phage DNA was extracted with the Wizard Genomic DNA Purification Kit (Promega). DNA restriction assays were performed according to standard NEB protocols and restriction fragments were assessed by agarose gel electrophoresis.

Type I-C CRISPR–Cas system assay and expression in *P. aeruginosa* PAO1

Type I-C CRISPR–Cas function was tested by electroporating a strain that naturally contains I-C Cas genes (strain F11) with pHERD30T plasmids encoding crRNAs that target phages. To express this system heterologously in PAO1, the four effector Cas genes (*cas3*, *cas5*, *cas8* and *cas7*) were cloned into pUC18T-lac and inserted in the PAO1 chromosome as described above. After removal of the gentamicin marker, this strain was electroporated with the same pHERD30T crRNA-encoding plasmids to confirm function after induction with IPTG and arabinose.

Type I-F CRISPR–Cas system expression in *P. aeruginosa* PAO1

To express the type I-F system heterologously in PAO1, all I-F Cas genes (*cas1*, *cas3*, *csy1*, *csy2*, *csy3* and *csy4*) were cloned into pMMBHE plasmids and transformed into PAO1. Subsequently, this strain was electroporated with the pHERD30T crRNA-encoding plasmids to confirm function after induction with IPTG and arabinose. To maintain pHERD30T and pMMBHE in the same strain of *P. aeruginosa*, double selection of 30 μg ml^{-1} gentamicin and 100 μg ml^{-1} carbenicillin was used.

crRNA cloning and expression

All crRNAs used here were cloned into established entry vectors in the pHERD30T background. After removing a pre-existing BsaI cut site in the vector by mutagenesis, a pseudo-CRISPR array (that is, repeat–spacer–repeat for types I, V or VI, or an sgRNA scaffold for type II) in which the spacing sequence possessed two inverted BsaI digest sites was then cloned into the vector, to facilitate scarless cloning of new spacers. Desired spacer sequences were chosen randomly across the phage genome (with the correct PAM sequence for the cognate CRISPR–Cas system) and ordered as two complementary oligonucleotides that generate sticky ends when annealed, to be cloned into the entry vector, which was digested with BsaI. Spacer oligonucleotides were treated with polynucleotide kinase (PNK), annealed and ligated into the entry vector.

*Spy*Cas9 and sgRNA expression in *P. aeruginosa*

The *S. pyogenes cas9* gene was cloned into a pUC18T-Ara integration vector and then inserted into the attTn7 locus of PAO1. An sgRNA scaffold was constructed based on a previous design[26] with internal BsaI cut sites to enable insertion of pre-annealed oligos for scarless sgRNA design. This sgRNA scaffold was amplified with primers p30T-gRNA_BsaI and p30T-gRNA_BsaI_rev. The resulting product was inserted into the pHERD30T vector by Gibson assembly after backbone (pJW1) amplification by inverse PCR with primers gRNA_BsaI-p30T and gRNA_BsaI-p30T-rev. The sgRNA scaffold was positioned into pJW1 so that after BsaI cleavage the spacer insert +1 position would conincide with the pBAD transcription start site (TSS) +1 position. The resulting plasmid, pJB1, was digested with BsaI (NEB), followed by ligation of indicated pre-annealed oligonucleotides. Supplementary Table 3 contains a complete list of all target sequences. The sequence of the sgRNA construct with BsaI site locations is shown in Supplementary Table 3.

Cas9 in vitro cleavage

Cas9-based cleavage of the phage genome in vitro was conducted with purified Cas9 protein (NEB, M0386S), and the cleavage reaction based on Cas9-gRNA-tracrRNA was then performed according to the manufacturer's instructions (NEB). Cas9 crRNAs (Supplementary Table 3) were ordered as Alt-R CRISPR-Cas9 crRNAs from IDT and used without further modification. The tracrRNA was amplified using primers tracrRNA-FOR and tracrRNA-REV from a plasmid (pBR62). The tracrRNA was produced through a T7 RNAP reaction using dsDNA that encodes the tracrRNA downstream of a T7 RNAP promoter. Cas9 protein (NEB) was combined with pre-annealed crRNA and tracRNA complex at a 1:1 molar ratio. For targeting KasI-liberated ΦKZ gDNA, 500 ng of ΦKZ DNA was co-incubated with Cas9 ribonucleoprotein (RNP), which cleaves at position 183,270, and KasI in NEB buffer 3.1 for 1 h. After stopping the reaction by proteinase K treatment, products were assessed by agarose gel electrophoresis. For Cas9 digestion of whole DNA, the reaction was performed at 37 °C for 4 h with 300 ng of ΦKZ or DMS3 gDNA and the products were assessed by agarose gel electrophoresis. Two Cas9 guides were selected that cleave at positions 158,649 and 168,715 of the ΦKZ genome to liberate a fragment of approximately 10 kb.

Cas12a and crRNA design for expression in *P. aeruginosa*

The humanized allele of the *cpf1* gene of *M. bovoculli* (MBO_03467, KDN25524.1) was sub-cloned from pTE4495 (Addgene) into pUC18T-lac using primers pUC_cpf1_F and pUC_cpf1_R and inserted in the PAO1 chromosome as described above. A Cpf1 repeat–spacer–repeat pseudo-CRISPR array was synthesized as oligonucleotides, annealed, ligated into a pHERD30T vector and digested with NcoI and HindIII. Spacer sequences were cloned into the resulting vector (pJB2) after digestion with BsaI and ligation of pre-annealed spacer oligonucleotide pairs.

Cas13a and crRNA design for expression in *P. aeruginosa*

The wild-type allele of the *cas13* gene of *L. seeligeri* and *Leptotrichia shahii* was sub-cloned from p2CT-His-MBP-Lse_C2c2_WT and p2CT-His-MBP-Lsh_C2c2_WT (Addgene) into pUC18T-lac. *Lse*Cas13 and *Lsh*Cas13 were inserted in the PAO1 chromosome as described above. An *Lse*- and

an *Lsh*Cas13a repeat–spacer–repeat pseudo-CRISPR array were synthesized as oligonucleotides, annealed, ligated into a pHERD30T vector and digested with NcoI and EcoRI. Spacer sequences were cloned into the resulting vectors (pSDM057 and pSDM070, respectively) after BsaI digestion and ligation of pre-annealed spacer oligonucleotide pairs. crRNA expression vectors were introduced into PAO1 tn7::*cas13*^{Lse} and PAO1 tn7::*cas13*^{Lsh}. The resulting strains were assayed for phage sensitivity under standard phage plating conditions, with induction of both Cas13 and the gRNAs (50 μg ml^{-1} gentamicin, 0.1% (L)-arabinose, 1 mM IPTG).

EcoRI expression in *P. aeruginosa*

The wild-type alleles of the M-EcoRI (methyltransferase) and R-EcoRI (endonuclease) genes were sub-cloned from pSB1A3 EcoRI Methylase-AmilCP and pSB1A3 EcoRI-RTX with EcoRI Methylase-AmilCP (Addgene plasmid 85166 and 85165), p2CT-His-MBP-Lse_C2c2_WT and p2CT-His-MBP-Lsh_C2c2_WT (Addgene) into pHERD30T using Gibson assembly. The resulting plasmids, pSDM160 and pSDM161, were electroporated into PAO1 Δ*hsdR* (SDM020). The resulting strains were assayed for phage sensitivity under standard phage plating conditions, with induction of both M-EcoRI (for genome protection) and R-EcoRI (50 μg ml^{-1} gentamicin, 0.1% (L)-arabinose).

Restriction-modification assay

PAO1, PAK and PAO1 Δ*hsdR* were grown to saturation in LB at 37 °C. Then, 4 ml of 0.7% molten top agar with 10 mM $MgSO_4$ were seeded with 100 μl saturated culture and spread on 20 ml 10 mM $MgSO_4$ LB agar plates. Tenfold serial dilutions of bacteriophage JBD30 and ΦKZ propagated on strain PAO1 and PAK were spotted (2.5 μl) on plates. Plates were incubated at 37 °C overnight and imaged the following day.

Chromosomal knockout of *hsdR*

To delete *hsdR* (NP_251422.1) from the PAO1 genome, a PAO1 strain that expresses Cas9 from *S. pyogenes* was programmed to express an sgRNA against *hsdR*: GCCCTCATCGAAGAAACCAG. In addition, the pHERD30T plasmid expressing this sgRNA was engineered to carry a repair template. This repair template consisted of the 500 bp upstream of *hsdR* and the 500 bp downstream of *hsdR* directly enjoined to one another. Induction of Cas9 and sgRNA led to cellular toxicity, as measured by $OD_{600\,nm}$ (data not shown). Survivors of Cas9-mediated targeting were isolated. The *hsdR* locus of survivors was then amplified by colony PCR with primers binding outside of the region encompassed by the repair template. Products were resolved by gel electrophoresis in a 1.0% agarose TAE gel at 100 V and visualized using the SYBR Safe DNA stain. Amplicons of a reduced size were then sequenced, which confirmed the chromosomal deletion of *hsdR*. A clone with the correct deletion, SDM020, was chosen for downstream experiments.

Construction of fusion proteins

Plasmids expressing Cherry alone, Cherry-Cas3, or Cherry-Cas8 (of type I-C and type I-F systems) and Cherry-TopA were constructed by Gibson assembly in the pHERD30T plasmid digested with SacI and PstI. These fusions have a GGAGGCGGTGGAGCC nucleotide (G-G-G-G-A amino acid) linker sequence in between them. *cherry* was amplified from SF-pSFFV-sfCherryFL1M3_TagBFP (provided by the laboratory of B. Huang, UCSF). *cas3* and *cas8* of the type I-C and type I-F systems were amplified from LL77 and PA14, respectively. *topA* was amplified from gDNA of PAO1. All of the primers that were used for the construction of the plasmids are listed in Supplementary Table 3.

Plasmids expressing Cherry–Cas9 (pESN28) (primers: prESN74, prESN75, prESN76, prESN77) and Cherry–Cas9–ORF152 (pESN29) (prESN80, prESN81, prESN82, prESN83) were constructed by Gibson assembly in the pHERD30T plasmid. Cherry–ORF152 (pESN32) (prESN91, prESN92) was constructed by PCR amplification to omit Cas9 and ligated using a KLD reaction. An SGAAAAGGSQK amino acid linker connects Cherry to the other proteins. A GGGGS amino acid linker connects Cas9 and ORF152. The plasmid expressing cMyc–ORF152 (prESN153, prESN154) was constructed by PCR amplification (primers: prESN153 and prESN154) using the p30-Cas9-ORF152 plasmid as the template, followed by a PCR clean-up and KLD ligation to circularize. This removed Cas9 from the plasmid, and added cMyc. All plasmids were designed using SnapGene.

To construct translational fusions of proteins, desired gene fragments including *cas9*, M-EcoRI, R-EcoRI, *orf152* (NP_803718.1) and sfCherry2 were amplified by PCR from templates pHERD30T::Cas9, pSB1A3::M.EcoRI-AmilCP, pSB1A3::M.EcoRI-AmilCP;R.EcoRI-RTX, ΦKZ gDNA and pHERD30T::sfCherry2, respectively. pHERD30T was linearized by restriction digestion with SacI-HF and PstI-HF or by PCR. PCR primers included 15–30-bp overhangs for Gibson assembly. Overhangs that enjoin genes included amino acid linkers GGGGS or GGSGGS. PCRs were treated with DpnI to eliminate template DNA. All PCRs and restriction digests were purified using a PCR clean-up kit (Zymo Research DNA Clean and Concentrator Kits). Linearized vector and gene fragments were treated with Gibson assembly, and Gibson assembly reactions were transformed into competent *E. coli*. Plasmid products were isolated by minipreparation and submitted for sequencing. Correct assemblies were electroporated into appropriate *P. aeruginosa* strains.

Isolation of Cas13 escapers

For identifying escapers of Cas13a RNA targeting, 3 μl of high-concentration ΦKZ lysates was mixed with 150 μl of overnight cultures of SDM078, SDM109 or SDM107 (strains expressing *Cas13a* and a gRNA against ΦKZ). After incubating at 37 °C for 10 min, samples were mixed with 4 ml of 0.7% agar, 10 mM $MgSO_4$, 1 mM IPTG and 0.1% arabinose and plated in LB agar plates with gentamicin (50 μg ml^{-1}). After overnight incubation, plates were examined for the presence of escaper plaques. Escapers were formed in SDM078 and SDM109, but not in SDM107. Ten phages that escaped targeting from SDM078 and SDM109 were purified and the protospacer locus was amplified using PCR and subsequently sequenced. Six unique outcomes were identified and shown in Fig. 4.

Growth curve experiments

Growth curve experiments were carried out in a Synergy H1 microplate reader (BioTek, with Gen5 software). Cells were diluted 1:100 from a saturated overnight culture with 10 mM $MgSO_4$ and antibiotics and inducers, as appropriate. Diluted culture (140 μl) was added together with 10 μl of phage to wells in a 96-well plate. This plate was cultured with maximum double orbital rotation at 37 °C for 24 h with $OD_{600\,nm}$ measurements every 5 minutes.

Immunofluorescence

Sample growth. Overnight cultures (5 ml) of a strain expressing Cas9 and an sgRNA targeting ΦKZ (SDM065), and a strain expressing cMyc–ORF152 (bESN27), were grown at 30 °C in LB medium with gentamicin. A 1:30 back dilution of the overnight culture into LB was grown at 30 °C for 1 h. Protein and guide expression was induced with 0.1% arabinose and 0.5 mM IPTG, respectively. After 1 h of expression, an aliquot of uninfected cells was fixed and the remaining cultures were infected with ΦKZ using a MOI of 1.5. Infected cell aliquots were collected and fixed at 60 min post-infection.

Fixation. This protocol was adapted from a previous study[27]. Samples were fixed with 5× fixing solution (12.5% paraformaldehyde, 150 mM KPO_4 pH 7.2) and incubated for 15 min at room temperature, followed by 20 min on ice. Samples were then washed in PBS three times and finally resuspended in GTE (50 mM glucose, 10 mM EDTA pH 8.0, 20 mM Tris-HCl pH 7.65) with 10 μg ml^{-1} lysozyme. Resuspended cells were transferred to polylysinated coverslips and dried. Once dry, the coverslips were incubated in cold methanol for 5 min and cold acetone for 5 min. Cells were rehydrated by a rinse in PBS, followed by a 3-min

incubation in PBS with 2% BSA blocking solution. Cells were incubated with a 1:50 dilution of primary antibody (Cas9 (7A9-3A3): sc-517386 or cMyc (9E10): sc-40) in PBS with 2% BSA for 1 h, followed by 3 washes of 7 min each in fresh PBS with 2% BSA. Coverslips were then incubated in the dark for 1 h with secondary antibody (goat anti-mouse Alexa Fluor 555, Life Technologies A-21424) diluted 1:500 in PBS with 2% BSA. DAPI was added for the final 10 min of the incubation. Cells were washed in PBS 3 times for 7 min. Coverslips were then placed on slides using mounting medium (v/v 90% glycerol, v/v 10% Tris pH 8.0 and w/v 0.5% propyl-gallate) and sealed with clear nail polish.

Microscopy and analysis. Images were collected using a Zeiss Axiovert 200M microscope, and were processed using NIS Analysis software at the UCSF Nikon Imaging Center. Compartments and cells were manually selected using the Simple ROI Editor. Background subtractions were conducted for each cell and compartment separately. The corrected intensities for the cytoplasmic and compartment regions were averaged over the area of each cell and compartment using Matlab and plotted using GraphPad Prism 6. Data from two pooled replicates were fitted with a line, with 95% confidence intervals shown as dashed lines. The slope is as reported in the plots.

Live-cell imaging

For live-cell imaging of ΦKZ infection, freshly grown cells from LB plates were picked and resuspended in 100 µl of LB medium. Around 5–10 µl of the samples was spotted on 0.85% agarose pads with 1:5 diluted LB, arabinose (0.01–0.05%) and DAPI (2 µg ml^{-1}). Samples were then incubated in a humidified chamber and allowed to grow for 3 h at 30 °C. ΦKZ lysate (5 µl)was spotted on top of the agar pad and the samples were grown for an additional 1 h. Cells were visualized in the microscope after being covered with a cover slip. A Nikon Ti2-E inverted microscope equipped with the Perfect Focus System (PFS) and a Photometrics Prime 95B 25-mm camera were used for live-cell imaging. Images were processed using NIS Elements AR software.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All data generated or analysed during this study are included in this published article and its Supplementary Information files.

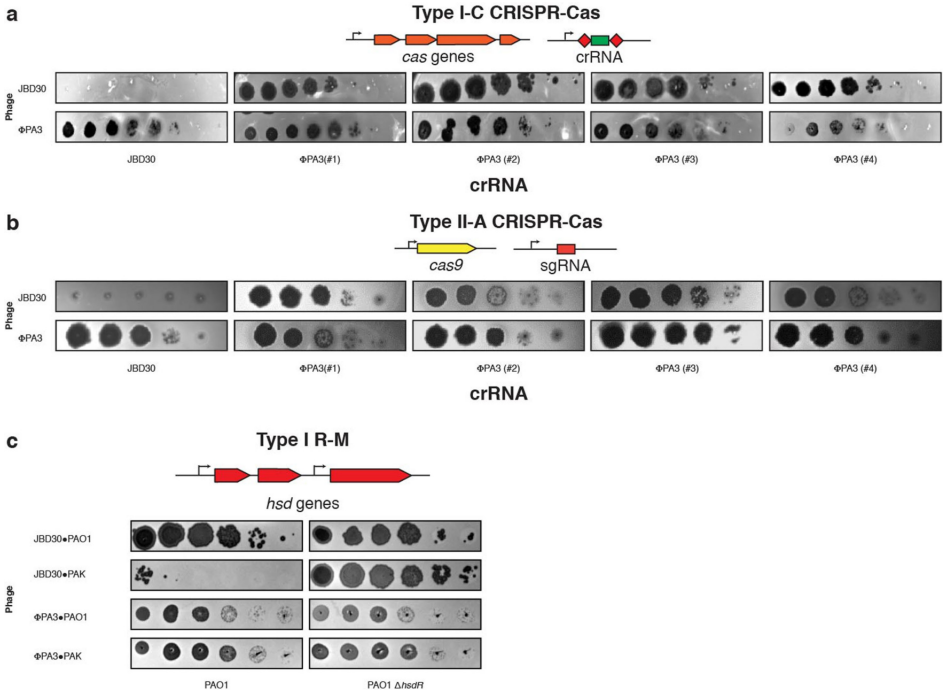
25. Choi, K.-H. & Schweizer, H. P. mini-Tn7 insertion in bacteria with single *att*Tn7 sites: example *Pseudomonas aeruginosa*. *Nat. Protoc.* **1**, 153–161 (2006).
26. Jinek, M. et al. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. *Science* **337**, 816–821 (2012).
27. Cowles, K. N. et al. The putative Poc complex controls two distinct *Pseudomonas aeruginosa* polar motility mechanisms. *Mol. Microbiol.* **90**, 923–938 (2013).

Acknowledgements Research in the J.B.-D. laboratory was supported by the University of California San Francisco Program for Breakthrough in Biomedical Research, which is funded in part by the Sandler Foundation, and an NIH Office of the Director Early Independence Award (DP5-OD021344 (J.B.-D.) and R01-GM127489 (J.B.-D.)). This work was also supported by HHMI (D.A.A.) and NIH grants R35GM118099 (D.A.A.) and GM104556 (D.A.A., J.P.). ΦKZ, JBD30, JBD68, D3 and F8 were provided by the laboratory of A. Davidson and ΦPA3 by the D.A.A. laboratory; phage DMS3m was a gift from G. O'Toole; the *S. pyogenes cas9* expression plasmid for integration in the PAO1 chromosome was from J. M. Peters and C. A. Gross; pTE4495 (MbCpf1/MbCas12a) (Addgene plasmid 80339) was a gift from E. Welker; *Lse*Cas13a (Addgene plasmid 83486) was a gift from J. Doudna; pSB1A3 EcoRI Methylase-AmilCP and pSB1A3 EcoRI-RTX with EcoRI Methylase-AmilCP (Addgene plasmids 85166 and 85165) were gifts from R. Dowell; and the type I-F crRNA expression plasmid pAB04 was from A. Borges (J.B.-D. laboratory).

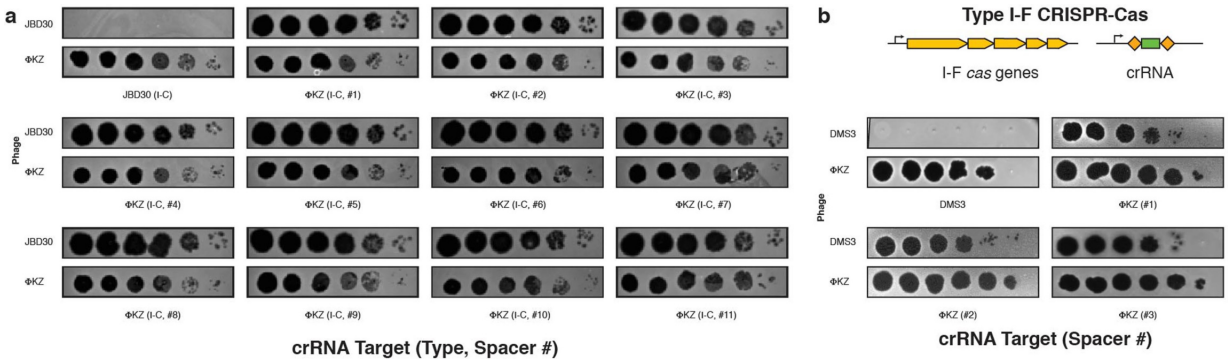
Author contributions S.D.M. constructed strains and performed experiments relating to restriction modification, relocalization of enzymes, Cas13, Cas12 and all liquid infection assays, and prepared figures. E.S.N. performed fluorescence microscopy experiments with guidance from V.C. and J.P., and strain construction and data analysis under the supervision of D.A.A. S.G. performed microscopy experiments, strain construction, plaque assays, data analysis, in vitro digestion assays and Cas13 escaper experiments. L.M.L. performed type I-C Cas3 experiments. J.D.B. constructed and performed experiments with Cas9-expressing strains and performed in vitro digestion assays. A.T. isolated and sequenced Cas13 escaper phages. J.B.-D. conceived the project, performed Cas3 and Cas9 experiments, supervised all experiments and wrote the manuscript together with S.D.M. All authors edited the manuscript.

Competing interests J.B.-D. is a scientific advisory board member of SNIPR Biome and Excision Biotherapeutics and a scientific advisory board member and co-founder of Acrigen Biosciences.

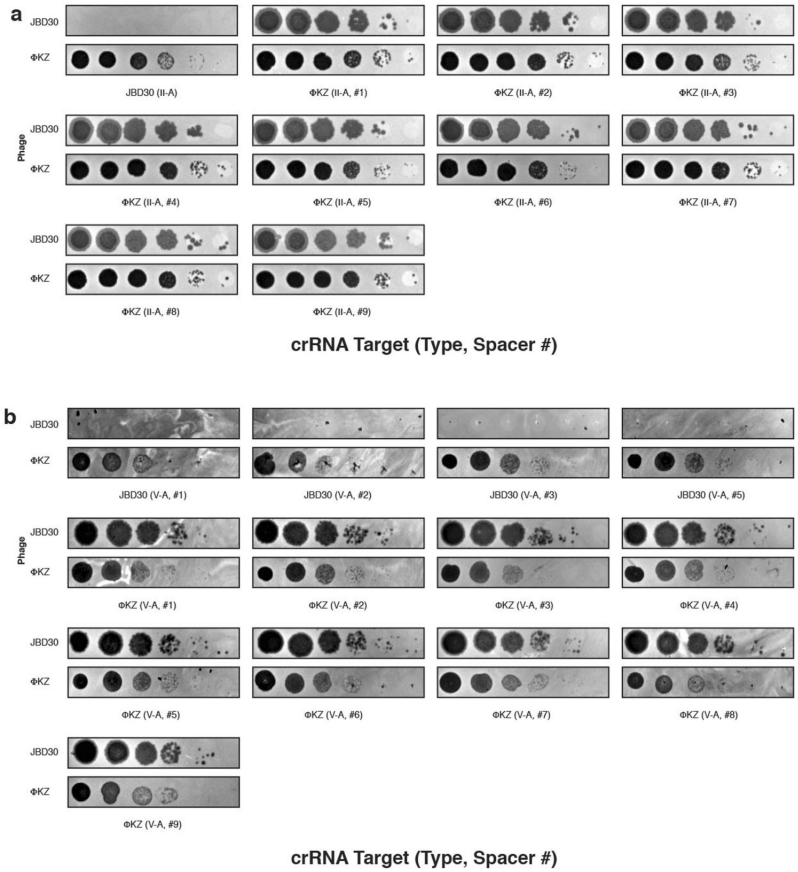
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1786-y.
Correspondence and requests for materials should be addressed to J.B.-D.
Peer review information *Nature* thanks Eugene Koonin and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



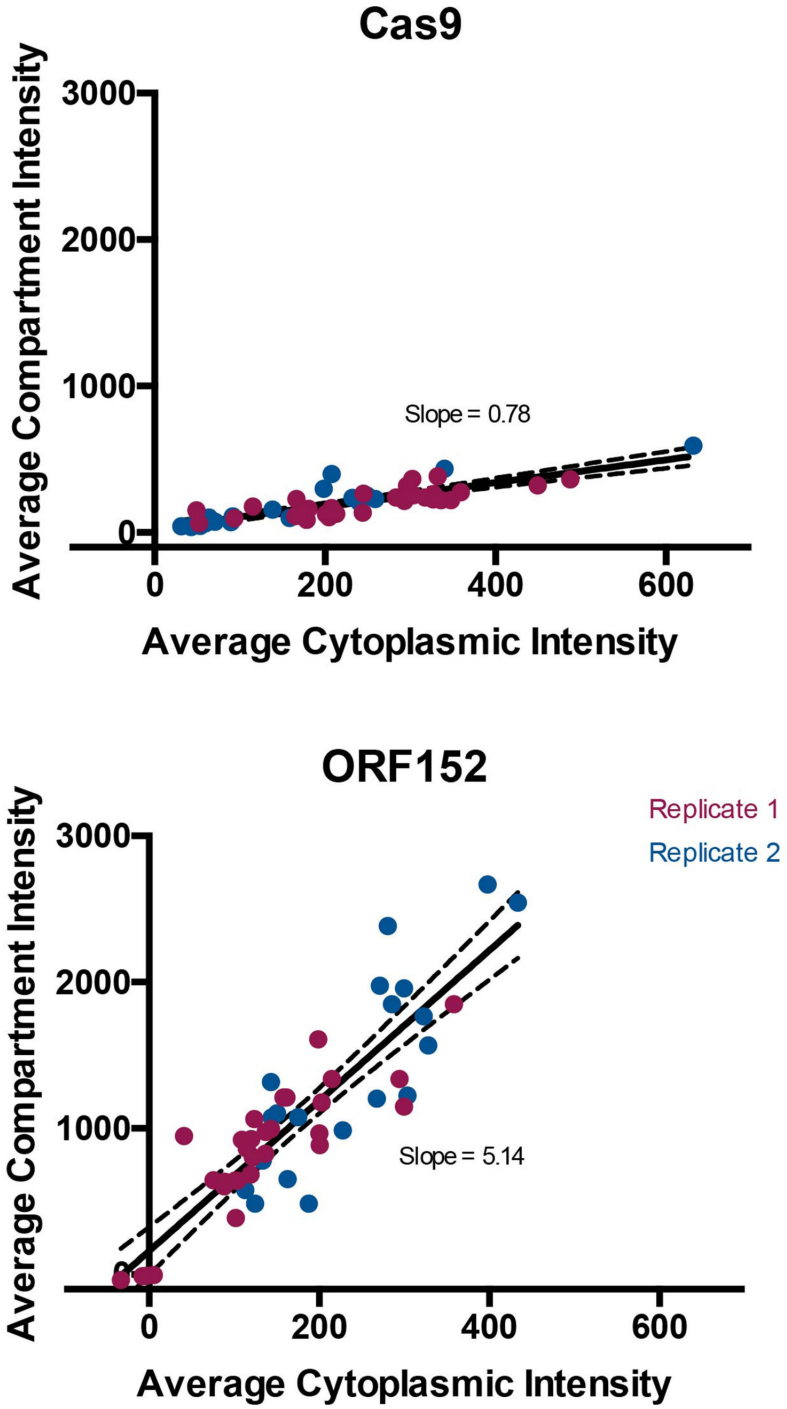
Extended Data Fig. 1 | Jumbo phage ΦPA3 resists targeting by CRISPR–Cas systems and an R-M system. a, Strain PAO1 was engineered to express the type I-C Cas genes and distinct crRNAs that target the indicated phages. **b**, Strain PAO1 was engineered to express the type II-A Cas9 protein and distinct sgRNAs that target the indicated phages. **c**, The endogenous type I R-M system (hsdRSM) in wild-type PAO1 and a mutant strain with an isogenic *hsdR* knockout (PAO1Δ *hsdR*) was assayed using phages that were propagated on PAO1 or PAK as indicated, and strains were subjected to a plaque assay. All plaque assays were conducted as in Fig. 1a and were replicated twice with similar results.



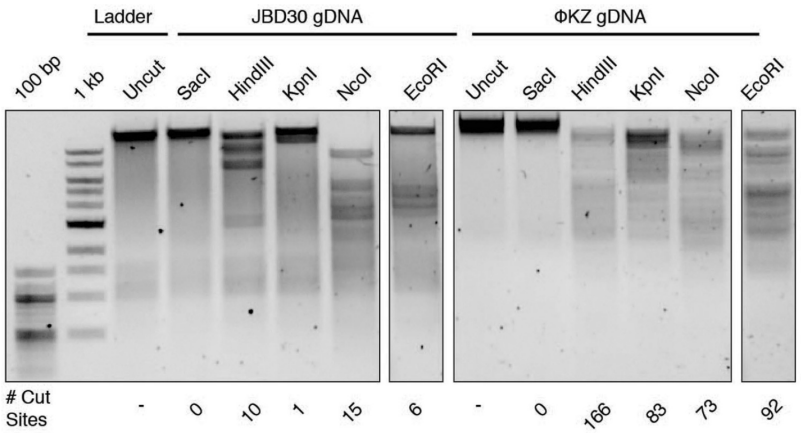
Extended Data Fig. 2 | Phage ΦKZ resists *P. aeruginosa* type I-C and type I-F CRISPR–Cas immunity. a, Strain PAO1 was engineered to express the type I-C Cas genes and distinct crRNAs that target phage JBD30 and phage ΦKZ. **b**, Strain PAO1 was engineered to express the type I-F Cas genes and distinct crRNAs that target phage JBD30 and phage ΦKZ. All plaque assays were conducted as in Fig. 1a and were replicated two or more times with similar results.



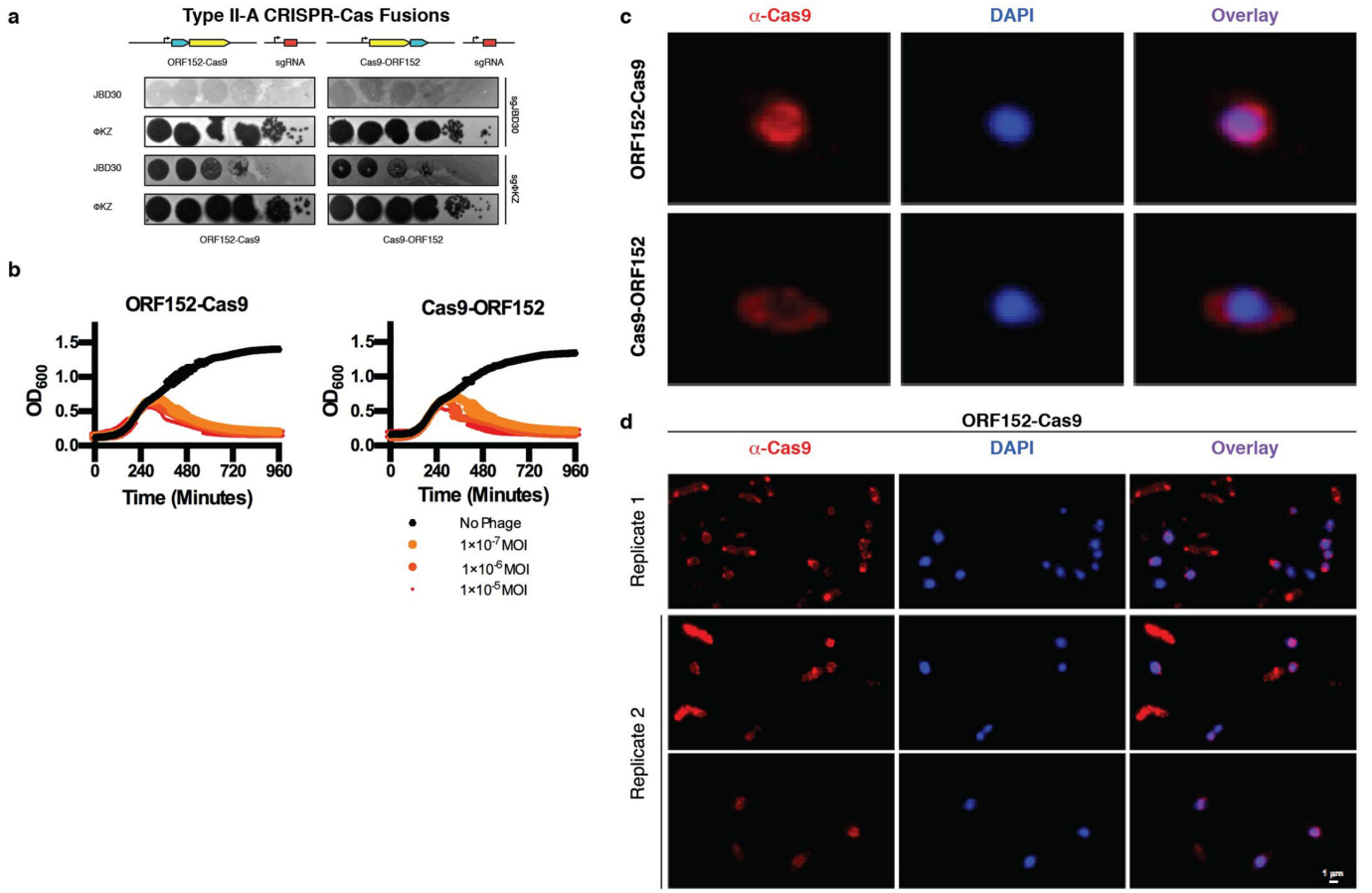
Extended Data Fig. 3 | Phage ΦKZ resists targeting by heterologous type II-A and V-A CRISPR–Cas systems. a, Strain PAO1 was engineered to express the type II-A Cas9 protein and distinct sgRNAs that target the indicated phages. **b**, Strain PAO1 was engineered to express the type V-A Cas12a protein and distinct crRNAs against the indicated phages. All plaque assays were conducted as in Fig. 1a and were replicated two or more times with similar results.



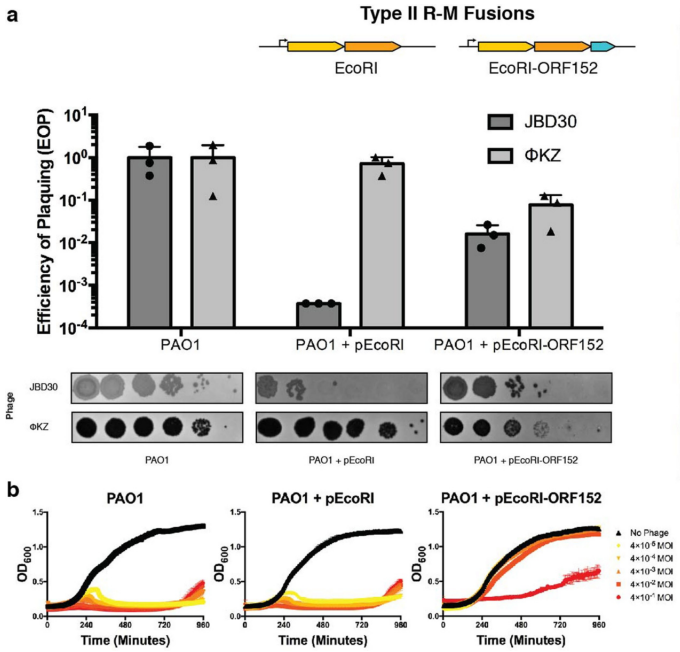
Extended Data Fig. 4 | Quantification of Cas9 localization during phage ΦKZ infection of *P. aeruginosa*. The localization of Cas9 and ORF152 in the cytoplasm and shell during ΦKZ infection was quantified. Data points (individual cells) from two pooled replicate experiments are fitted with a line and 95% confidence intervals are indicated with dashed lines. The slope is as reported in the plots.

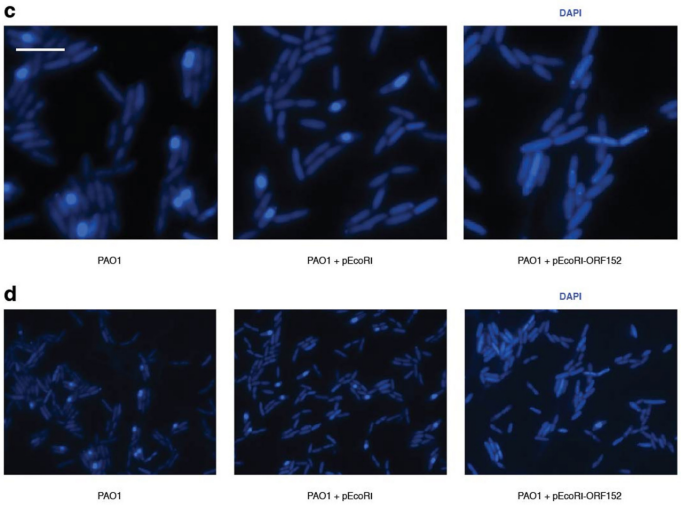


Extended Data Fig. 5 | Phage ΦKZ gDNA is susceptible in vitro to cleavage by restriction endonucleases. gDNA was isolated from phages ΦKZ and JBD30 and subjected to digestion with the indicated restriction enzymes in vitro. The number of cut sites for each enzyme is shown at the bottom of the gels. Products were visualized on a 0.7% agarose gel stained with SYBR Safe. The experiment was replicated twice with similar result.

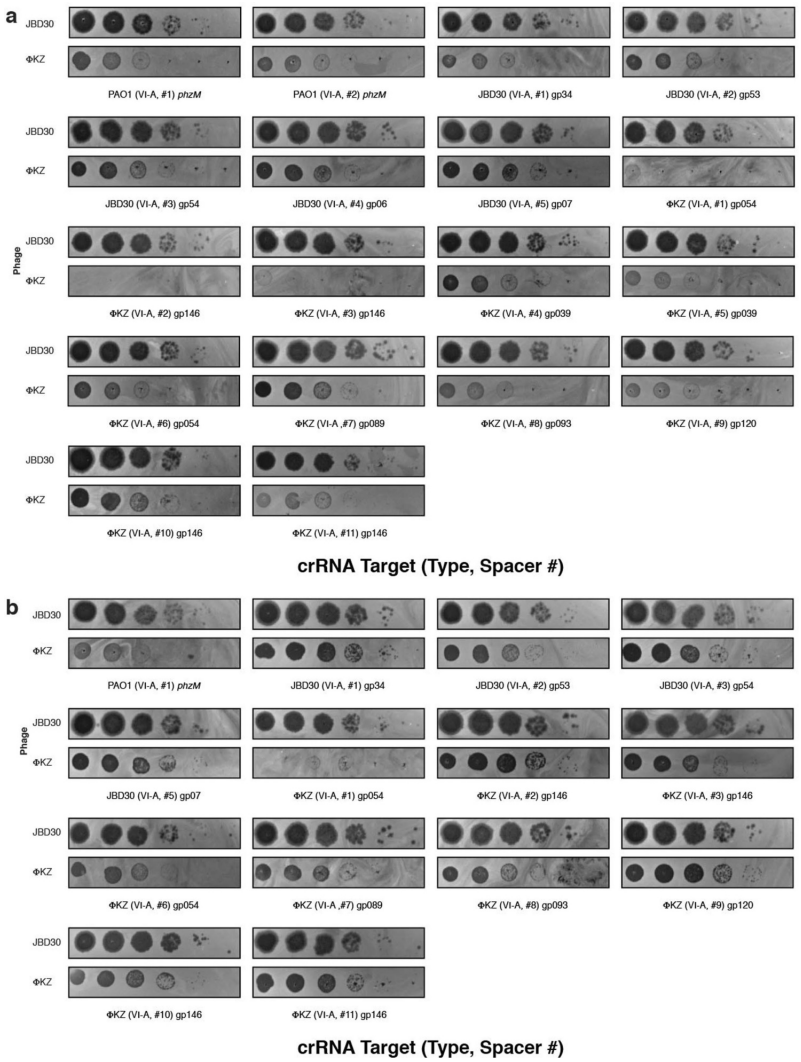


Extended Data Fig. 6 | Cas9 protein fused to ORF152 localizes to the periphery of the shell, but does not enable immune activity against ΦKZ. **a**, Strain PAO1 was engineered to express Cas9 fused to ORF152 at either the N or the C terminus, with sgRNAs targeting the indicated phages. **b**, Growth curves measuring the $OD_{600\,nm}$ of PAO1 cells that were infected with ΦKZ at the indicated MOI. **c**, Fluorescence microscopy of PAO1 fusion strains, immunostained for Cas9, in cells that express ORF152–Cas9 or Cas9–ORF152. **d**, Fluorescence microscopy of *P. aeruginosa*, immunostained for Cas9, in cells that express ORF152–Cas9. The DAPI stain in **c**, **d** shows the phage DNA within the shell. All plaque assays were conducted as in Fig. 1a and were replicated four times with similar results. The growth curve experiments (**b**) were replicated three times with similar results. Microscopy (**c**, **d**) was replicated twice with similar results.

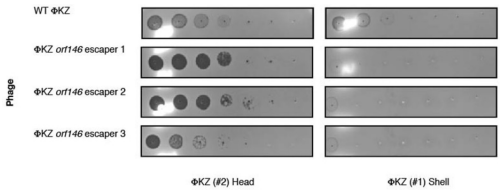




Extended Data Fig. 7 | Fusion of EcoRI restriction enzymes to ORF152 enables immune activity. a, Strain PAO1 was engineered to express EcoRI or an EcoRI–ORF152 fusion protein. Plaque assays were conducted as in Fig. 1a and quantified ($n = 3$). **b**, Growth curves measuring the $OD_{600\,nm}$ of PAO1 cells that were infected with ΦKZ at the indicated MOI. **c**, **d**, Live-cell fluorescence imaging (with a wide field of view in **d**) of *P. aeruginosa* strains that were engineered to express EcoRI or EcoRI–ORF152. The DAPI stain shows the phage DNA. All plaque assays were replicated three times with similar results. The growth curve experiments (**b**) were replicated twice with similar results. Microscopy (**c**, **d**) was replicated twice with similar results. Scale bar, 5 µm.



Extended Data Fig. 8 | Phage ΦKZ is sensitive to RNA targeting with Cas13. **a**, Strain PAO1 expressing *Lse*Cas13a and distinct crRNAs that target the indicated genes. **b**, Strain PAO1 expressing *Lsh*Cas13a and distinct crRNAs that target the indicated genes. All plaque assays were conducted as in Fig. 1a and were replicated twice with similar results.



Extended Data Fig. 9 | Phage ΦKZ escaper mutants are selected by Cas13-mediated RNA targeting. Strain PAO1 expressing *Lse*Cas13a and a crRNA that targets the indicated gene. Plaque assays were conducted as in Fig. 1a using wild-type and escaper-mutant ΦKZ. Wild-type ΦKZ is targeted by both strains and the faint clearings observed here are not observed as plaques in a full-plate assay. All plaque assays were replicated twice with similar results.



Extended Data Fig. 10 | Observation of DAPI-stained phage DNA adjacent to a nascent shell that contains Cherry–TopA. Live-cell fluorescence imaging of *P. aeruginosa* strains that were engineered to express Cherry–TopA and infected with ΦKZ. The DAPI stain labels DNA. Microscopy was replicated three times with similar results.

nature research

Corresponding author(s): Joseph Bondy-Denomy

Last updated by author(s): Sep 27, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Pictures of plaque assays and gels were taken using Image Lab (TM) Version 6.0.1 (C) 2017, Bio-Rad Laboratories, Inc. Growth curve assays were collected using BioTek Synergy H1 software Gen5 3.05.11. Microscopy was collected using NIS Elements-AR software 5.0.2.00.
Data analysis	Pictures of plaque assays and gels were edited using Image Lab (TM) Version 6.0.1 (C) 2017, Bio-Rad Laboratories, Inc. and ImageJ 1.50i. Plate reader growth curves were plotted using Prism 6. Plaque assay quantification was plotted using Prism 6. For Cas9 exclusion quantification, cells and compartments were chosen using Simple ROI Editor, from NIS Elements-AR 5.0.2.00. Background subtraction and average intensities were performed using MatLAB R2018b 9.5.0.944444 and plotted using Prism 6.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Included in main text.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	For I-C, I-F, II-A, V-A, and VI-A CRISPR and type I and type II restriction enzyme plaque assays, n>=2 as output of targeting was either full plaque formation or complete protection and thus was not quantified. For II-A and VI-A CRISPR growth curve assays, n=3. For EcoRI Fusion plaque assays, n=3. For EcoRI Fusion growth curve assay, n=3. For microscopy, one sample was imaged and representative cells are demonstrated. For immunfluorescence quantification, two replicates were imaged and selected cells were chosen (n=30 for Replicate 1, n=20 for Replicate 2) and analyzed.
Data exclusions	Gels in Figure 3 and Extended Figure 5 were cropped for clarity and conciseness. Full gels are provided in a .pdf.
Replication	Plaque assays and growth curves were all performed at least twice, showing replication each time. Here we show representative examples. Microscopy data shown is representative examples of phenotypes observed throughout the sample(s).
Randomization	Samples were organized into experimental groups depending on genotype, including genes and crRNAs encoded by the bacterial strain. Randomization was not performed in this study.
Blinding	Blinding was not performed in this study, as it was not relevant given that all results are quantitative, or represent binary states (e.g. exclusion vs. inclusion).

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Santa Cruz Biotechnology, Inc. Cas9 (7A9-3A3): sc-517386, Santa Cruz Biotechnology, Inc cMyc Antibody (9E10): sc-40
Validation	Cas9, from the manufacturer's website: "Yes, this mouse monoclonal antibody, Cas9 (7A9-3A3): sc-517386, has been validated to work for immunofluorescent staining applications. Unfortunately, we do not yet have any published IF data for this clone on our website. However, this clone, (7A9-3A3) has been widely published in the literature for use in a variety of applications, including IF." cMyc: There are 5553 citations of this antibody from Santa Cruz in the literature and several examples of both western blots and IF on the product website. But they do not specifically address validation of it anywhere.

Article

IL-17a promotes sociability in mouse models of neurodevelopmental disorders

https://doi.org/10.1038/s41586-019-1843-6

Received: 17 December 2018

Accepted: 30 October 2019

Published online: 18 December 2019

Michael Douglas Reed[1,2,7], Yeong Shin Yim[1,2,7], Ralf D. Wimmer[1,2,3], Hyunju Kim[4], Changhyeon Ryu[1,2], Gwyneth Margaret Welch[1,2], Matias Andina[1,2], Hunter Oren King[2], Ari Waisman[5], Michael M. Halassa[1,2,3], Jun R. Huh[4,6]* & Gloria B. Choi[1,2]*

A subset of children with autism spectrum disorder appear to show an improvement in their behavioural symptoms during the course of a fever, a sign of systemic inflammation[1,2]. Here we elucidate the molecular and neural mechanisms that underlie the beneficial effects of inflammation on social behaviour deficits in mice. We compared an environmental model of neurodevelopmental disorders in which mice were exposed to maternal immune activation (MIA) during embryogenesis[3,4] with mouse models that are genetically deficient for contactin-associated protein-like 2 (*Cntnap2*)[5], fragile X mental retardation-1 (*Fmr1*)[6] or Sh3 and multiple ankyrin repeat domains 3 (*Shank3*)[7]. We establish that the social behaviour deficits in offspring exposed to MIA can be temporarily rescued by the inflammatory response elicited by the administration of lipopolysaccharide (LPS). This behavioural rescue was accompanied by a reduction in neuronal activity in the primary somatosensory cortex dysgranular zone (S1DZ), the hyperactivity of which was previously implicated in the manifestation of behavioural phenotypes associated with offspring exposed to MIA[8]. By contrast, we did not observe an LPS-induced rescue of social deficits in the monogenic models. We demonstrate that the differences in responsiveness to the LPS treatment between the MIA and the monogenic models emerge from differences in the levels of cytokine production. LPS treatment in monogenic mutant mice did not induce amounts of interleukin-17a (IL-17a) comparable to those induced in MIA offspring; bypassing this difference by directly delivering IL-17a into S1DZ was sufficient to promote sociability in monogenic mutant mice as well as in MIA offspring. Conversely, abrogating the expression of IL-17 receptor subunit a (IL-17Ra) in the neurons of the S1DZ eliminated the ability of LPS to reverse the sociability phenotypes in MIA offspring. Our data support a neuroimmune mechanism that underlies neurodevelopmental disorders in which the production of IL-17a during inflammation can ameliorate the expression of social behaviour deficits by directly affecting neuronal activity in the central nervous system.

The beneficial effects of infection and the ensuing inflammation on neurological disorders have previously been noted[9]. For example, a subset of children with autism spectrum disorder (ASD) exhibit temporary but considerable improvements of their behavioural symptoms during episodes of fever, a sign of systemic inflammation[1,2]. However, a mechanistic understanding of how fever-associated immune responses translate into behavioural relief—both at the molecular and neural level—is lacking. The aetiology of ASD includes both environmental and genetic risk factors[10]. Mouse models for ASD, in which mice contain mutations in ASD risk genes (including *Cntnap2*[5], *Fmr1*[6], and *Shank3*[7]), show behavioural abnormalities—including deficits in social interaction[11,12]. Similar behavioural abnormalities are observed in mouse models for environmental risk factors for developing ASD, such as exposure to maternal inflammation[13]. We sought to explore the mechanisms that enable the fever-associated rescue of sociability deficits, using both genetic and environmental mouse models for neurodevelopmental disorders.

LPS rescues sociability in MIA offspring

A febrile response can be exogenously induced by injecting mice with a low dose of LPS[14]. Indeed, the intraperitoneal administration of

[1]The Picower Institute for Learning and Memory, Massachusetts Institute of Technology, Cambridge, MA, USA. [2]Department of Brain and Cognitive Sciences, Massachusetts Institute of Technology, Cambridge, MA, USA. [3]McGovern Institute for Brain Research, Massachusetts Institute of Technology, Cambridge, MA, USA. [4]Department of Immunology, Blavatnik Institute, Harvard Medical School, Boston, MA, USA. [5]Institute for Molecular Medicine, University Medical Center of the Johannes Gutenberg-University Mainz, Mainz, Germany. [6]Evergrande Center for Immunologic Diseases, Harvard Medical School and Brigham and Women's Hospital, Boston, MA, USA. [7]These authors contributed equally: Michael Douglas Reed, Yeong Shin Yim.
*e-mail: Jun_Huh@hms.harvard.edu; gbchoi@mit.edu

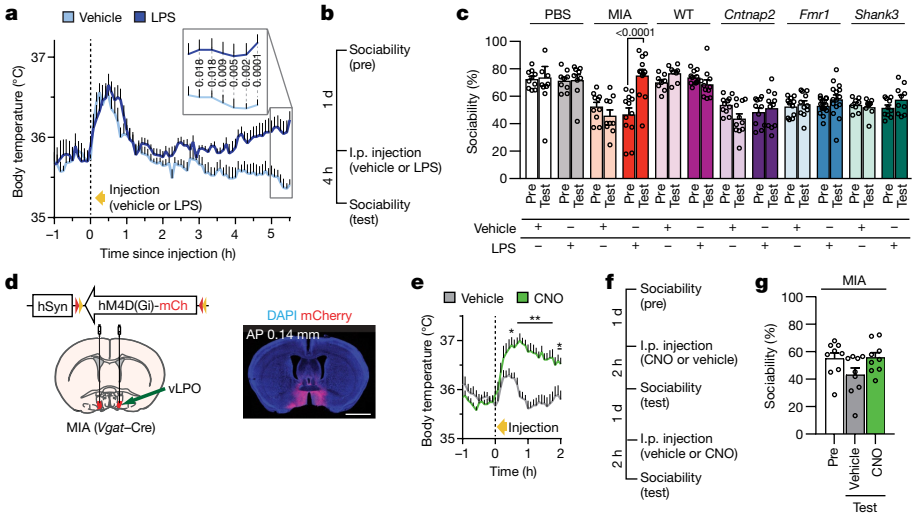


Fig. 1 | Immune stimulation rescues sociability deficits in MIA offspring. **a**, Body temperature profile after injection of saline (vehicle) or LPS in offspring of mothers injected with PBS (PBS offspring) (vehicle, $n = 11$ and LPS, $n = 11$ mice, from 5 independent experiments). The initial spike in body temperature is due to handling stress. **b**, **c**, Mice were tested for sociability (percentage of time spent investigating social object/total time spent investigating both social and inanimate objects) one day before LPS injection (pre-test; pre). Mice were then tested for sociability four hours after injection with vehicle or LPS (test). PBS offspring + vehicle, $n = 10$; PBS offspring + LPS, $n = 9$; MIA offspring + vehicle, $n = 10$; MIA offspring + LPS, $n = 12$; wild type (WT) + vehicle, $n = 8$; wild type + LPS, $n = 11$; *Cntnap2* mutant + vehicle, $n = 11$; *Cntnap2* mutant + LPS, $n = 11$; *Fmr1* mutant + vehicle, $n = 11$; *Fmr1* mutant + LPS, $n = 15$; *Shank3* mutant + vehicle, $n = 8$; *Shank3* mutant + LPS, $n = 10$; from 3 independent experiments. I.p., intraperitoneal. **d**, Virus encoding the inhibitory DREADD (AAV_2-hSyn-DIO-hM4D(Gi)-mCherry) was targeted to the ventral part of the lateral preoptic nucleus (vLPO) of *Vgat*–Cre MIA mice. Scale bar, 2 mm. **e**, Body temperature profile after injection with vehicle or CNO. **f**, **g**, Mice were tested for sociability one day before injection (pre). On the following two days, mice received counterbalanced injections of either vehicle or CNO. Sociability was assessed two hours after injection. For experiments in **d**–**g**, $n = 9$ for all groups, from 2 independent experiments. All n values refer to the number of mice used. $^*P < 0.05$, $^{**}P < 0.01$, calculated by two-way repeated-measures analysis of variance (ANOVA) with Bonferroni's (**a**, **e**) or Sidak's (**c**) post-hoc tests, or one-way repeated-measures ANOVA with Tukey's post-hoc test (**g**). Graphs are mean ± s.e.m.

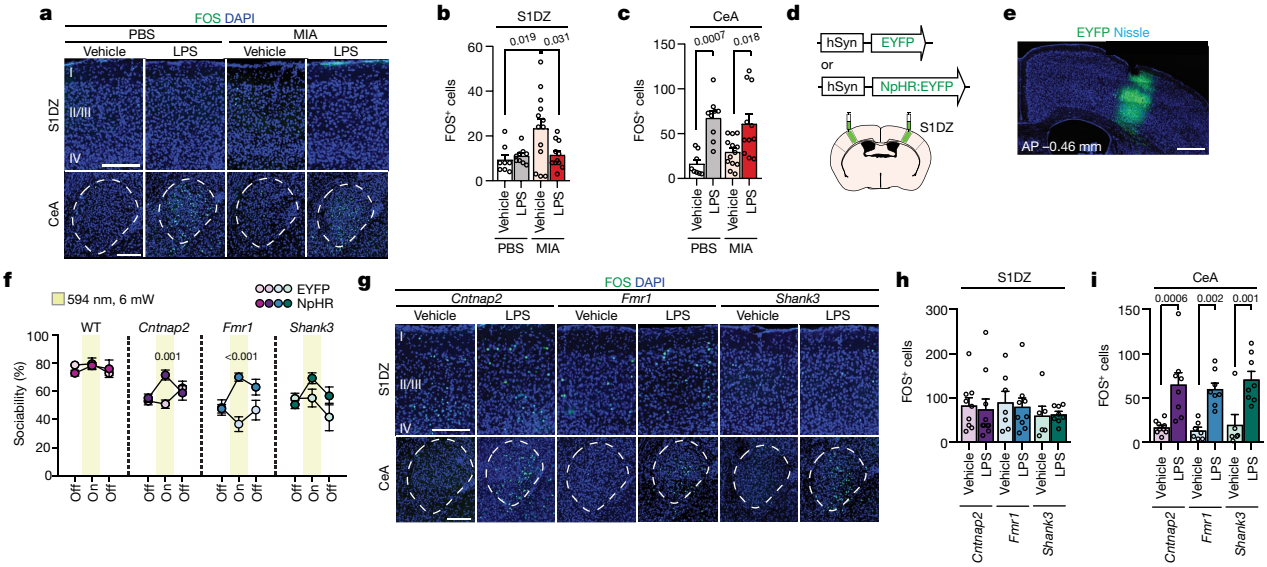


Fig. 2 | Immune stimulation reduces hyperactivation in the S1DZ of MIA offspring. **a**, Representative images illustrating FOS (green) expression in the S1DZ and CeA, after injection with vehicle or LPS. Scale bars, 200 µm. Numerals indicate cortical layers. **b**, **c**, Quantification of FOS-expressing cells in the S1DZ (**b**) and CeA (**c**). For experiments in **a**–**c**, PBS offspring + vehicle, $n = 8$; PBS offspring + LPS, $n = 9$; MIA offspring + vehicle, $n = 13$; MIA offspring + LPS, $n = 11$; from 3 independent experiments. **d**, **e**, AAV encoding either EYFP or EYFP fused to enhanced halorhodopsin (NpHR) was bilaterally injected into the S1DZ of monogenic mutant mice. Scale bar, 500 µm. **f**, Performance on sociability was assessed in the presence (on) and absence (off) of optical inhibition. For experiments in **d**–**f**, wild type + EYFP, $n = 7$; wild type + NpHR, $n = 8$; *Cntnap2* mutant + EYFP, $n = 11$; *Cntnap2* mutant + NpHR, $n = 9$; *Fmr1* mutant + EYFP, $n = 8$; *Fmr1* mutant + NpHR, $n = 12$; *Shank3* mutant + EYFP, $n = 8$; *Shank3* mutant + NpHR, $n = 10$; from 6 independent experiments. **g**, Representative images illustrating FOS expression in the S1DZ and CeA after injection of vehicle or LPS in monogenic mutant mice. Scale bars, 200 µm. **h**, **i**, Quantification of FOS-expressing cells after injection of vehicle or LPS in the S1DZ (**h**) and CeA (**i**). For experiments in **g**–**i**, *Cntnap2* mutant + vehicle, $n = 9$; *Cntnap2* mutant + LPS, $n = 10$; *Fmr1* mutant + vehicle, $n = 7$; *Fmr1* mutant + LPS, $n = 9$; *Shank3* mutant + vehicle, $n = 6$; *Shank3* mutant + LPS, $n = 8$; from 3 independent experiments. All n values refer to the number of mice used. Statistics calculated by two-way ANOVA with Tukey's post-hoc test (**b**, **c**) or Sidak's post-hoc test (**h**, **i**), or two-way repeated-measures ANOVA with Sidak's post-hoc test (**f**). Graphs are mean ± s.e.m.

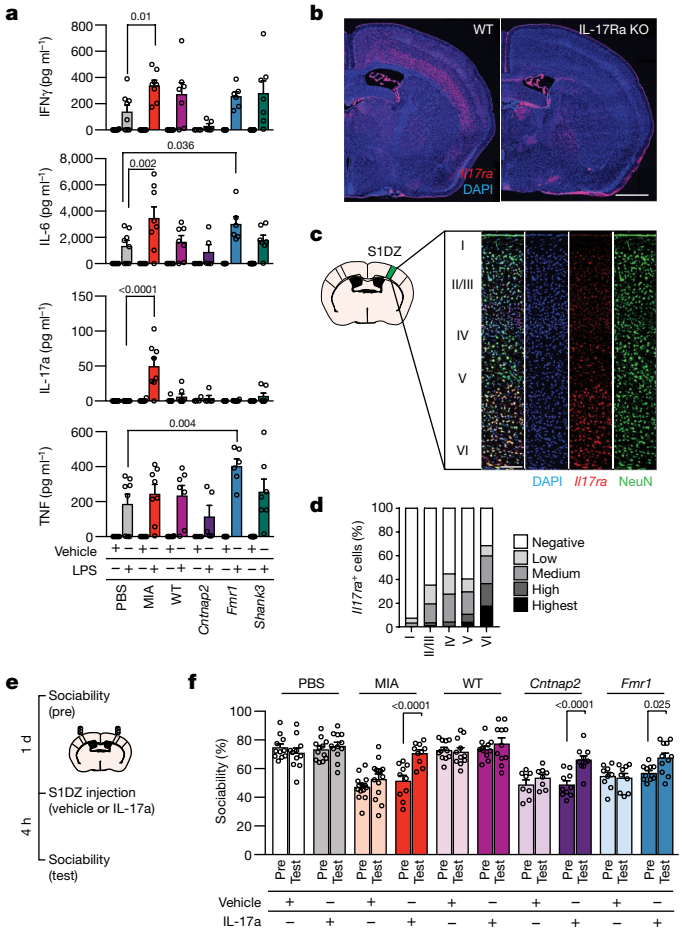


Fig. 3 | IL-17a rescues sociability deficits in both MIA offspring and monogenic mutant mice. a, IFNγ, IL-6, IL-17a and TNF levels in plasma after injection of vehicle or LPS. PBS offspring + vehicle, $n = 6$; PBS offspring + LPS, $n = 8$; MIA offspring + vehicle, $n = 6$; MIA offspring + LPS, $n = 8$; wild type + vehicle, $n = 6$; wild type + LPS, $n = 7$; *Cntnap2* mutant + vehicle, $n = 4$; *Cntnap2* mutant + LPS, $n = 5$; *Fmr1* mutant + vehicle, $n = 5$; *Fmr1* mutant + LPS, $n = 6$; *Shank3* mutant + vehicle, $n = 7$; *Shank3* mutant + LPS, $n = 7$; from 3 independent experiments. **b**, *Il17ra* expression in wild-type and IL-17Ra-knockout (KO) mice at anterior–posterior (AP) −0.58 mm, from 1 independent experiment. Scale bar, 1 mm. **c**, Colabelling of *Il17ra* and NeuN in the S1DZ, from 2 independent experiments. Scale bar, 100 µm. **d**, Quantification of *Il17ra* expression within the S1DZ according to cortical layer in PBS offspring. $n = 8$, from 2 independent experiments. **e**, **f**, Mice were tested for sociability one day before injection (pre) and after bilateral administration of vehicle or IL-17a into the S1DZ (test). PBS offspring + vehicle, $n = 11$; PBS offspring + IL-17a, $n = 12$; MIA offspring + vehicle, $n = 14$; MIA offspring + IL-17a, $n = 10$; wild type + vehicle, $n = 11$; wild type + IL-17a, $n = 11$; *Cntnap2* mutant + vehicle, $n = 8$; *Cntnap2* mutant + IL-17a, $n = 10$; *Fmr1* mutant + vehicle, $n = 9$; *Fmr1* mutant + IL-17a, $n = 11$; from 6 independent experiments. All n values refer to the number of mice used. Statistics calculated by two-way ANOVA with Dunnett's post-hoc test (**a**) or two-way repeated-measures ANOVA with Sidak's post-hoc test (**f**). Graphs are mean ± s.e.m.

LPS (50 µg kg^{-1}) to control mice—offspring born to mothers injected with PBS—led to a significant increase in body temperature (of about 0.5–1.0 °C) about 4–5 h after the treatment (Fig. 1a). To investigate whether LPS injection can ameliorate sociability deficits, we compared sociability in adult male mice one day before LPS injection (pre-test) and four hours after injection (test), at which point the increase in body temperature initiates (Fig. 1b). As previously reported[4,8,15,16], MIA offspring born to mothers injected with polyinosinic:polycytidylic acid (poly(I:C)) at embryonic day (E)12.5 exhibited impaired social approach behaviour during the pre-test (Fig. 1c, Extended Data Fig. 2). The *Cntnap2*, *Fmr1* and *Shank3* monogenic models also displayed sociability deficits, albeit with marked variability during the pre-test (Extended Data Fig. 1). We therefore compared MIA offspring to monogenic mutant mice that exhibited a sociability index lower than about 62%, the mean value for the three monogenic lines (Extended Data Fig. 1a).

LPS injections robustly rescued the characteristic deficits in sociability of MIA offspring; treated MIA offspring displayed an interest in social objects that was indistinguishable from that of controls (Fig. 1c, Extended Data Fig. 2). LPS injection did not affect baseline sociability in control offspring. The LPS-induced rescue of sociability was absent 72 h after the treatment, paralleling the transient nature of fever-associated improvement that has previously been observed in children with ASD[1] (Extended Data Fig. 3a–e). Furthermore, the LPS-induced rescue was observed in both young (2–5 months old) and aged (9–12 months old) mice (Fig. 1c, Extended Data Fig. 3f–j), which suggests that this is a generalizable phenomenon that is not confined to early adulthood. The LPS-induced rescue of the sociability deficits was also evident when assayed using a reciprocal social interaction test (Extended Data Fig. 3k). Finally, the LPS treatment of MIA offspring also suppressed the enhanced marble burying behaviour (Extended Data Fig. 3l), which shows that inflammation can relieve several of the behavioural phenotypes associated with MIA. Unlike in MIA offspring however, LPS treatment did not restore sociability in mutant mice that are deficient for *Cntnap2*, *Fmr1* or *Shank3* (Fig. 1c, Extended Data Fig. 2), indicating that rescue by LPS-driven inflammation may be applicable to only a subset of mouse models for neurodevelopmental disorders.

Acute fever does not restore sociability

We next probed whether fever has a role in the observed behavioural rescue. Unlike in control offspring, LPS treatment did not induce changes in body temperature in MIA mice (Extended Data Fig. 4a), which suggests that the febrile response might not be the main factor that contributes to the rescue. To directly test whether fever is dispensable, we sought to increase the body temperature of the mice without inducing systemic inflammation by targeting inhibitory designer receptors exclusively activated by designer drugs (DREADDs)[17] to GABAergic neurons in the ventral part of the lateral preoptic nucleus of *Vgat*–Cre (Methods) mice[18] (Fig. 1d). As previously reported[19], inhibition of these neurons led to an increase in body temperature of about 1 °C (Fig. 1e). However, the induction of febrile response alone did not promote social preference in MIA offspring (Fig. 1f, g, Extended Data Fig. 4b–e), confirming that fever per se is not the main driver of the LPS-induced rescue. Of note, *Vgat*–Cre MIA offspring exhibited sociability deficits comparable to those of wild-type MIA offspring, were unaffected by treatment with clozapine *N*-oxide (CNO) and showed an increase in sociability after treatment with LPS (Extended Data Fig. 4f–j).

LPS reduces FOS induction in the S1DZ

We previously established that adult MIA offspring display cortical abnormalities that are preferentially localized in the S1DZ[8], a subregion of the primary somatosensory cortex (S1) that is cyto-architecturally defined by the absence of a discernible fourth layer (Extended Data Fig. 5). The cortical phenotype is characterized by an overall increase in neural activity that, when reduced, can acutely rescue MIA-induced deficits in social behaviours[8]. We therefore investigated whether the LPS-induced behavioural rescue in MIA offspring is accompanied by changes in neural activity in the S1DZ. MIA offspring exhibited an increase in the number of cells in the S1DZ that express FOS (a marker for neuronal activation), relative to control offspring. However, in LPS-treated MIA offspring, the number of FOS$^+$ neurons in the S1DZ was reduced to the level of control offspring (Fig. 2a, b, Extended Data Fig. 6a–c). LPS injections did not elicit a generalized, brain-wide effect

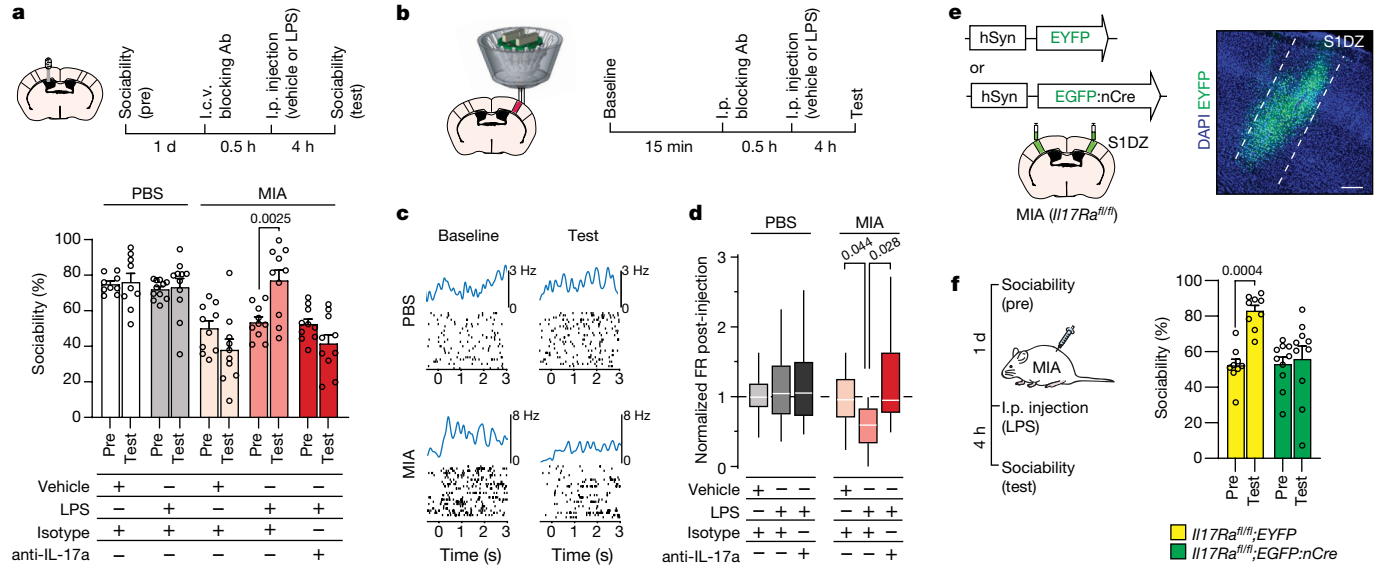


Fig. 4 | IL-17a is necessary for the LPS-induced rescue of sociability deficits and reduction of neural activity in the S1DZ in MIA offspring. a, Mice were tested for sociability one day before injection (pre). The following day, blocking antibody against IL-17a (anti-IL-17a) or isotype-control antibody (isotype) was administered intracerebroventricularly (i.c.v.), 30 min before injection of vehicle or LPS. Sociability was assessed four hours after injection of vehicle or LPS (test). PBS offspring + vehicle + isotype, $n = 9$; PBS offspring + LPS + isotype, $n = 11$; MIA offspring + vehicle + isotype, $n = 10$; MIA offspring + LPS + isotype, $n = 10$; MIA offspring + LPS + anti-IL-17a, $n = 10$; from 7 independent experiments. **b**–**d**, Firing rate of neurons of the S1DZ before and four hours after injection of vehicle or LPS, in PBS and MIA offspring pre-treated with isotype-control antibody or blocking antibody against IL-17a. **c**, Example raster plot with firing rate profile before (baseline) and after (test) treatment with LPS, from an PBS mouse and MIA mouse pre-treated with isotype-control antibody. **d**, Normalized firing-rate (FR) change after treatment represented as box-and-whisker plots, indicating median, interquartile range and data limits as defined by Tukey. PBS offspring + vehicle + isotype, $n = 65$ cells; PBS offspring + LPS + isotype, $n = 42$ cells; PBS offspring + LPS + anti-IL-17a, $n = 40$ cells; MIA offspring + vehicle + isotype, $n = 75$ cells; MIA offspring + LPS + isotype, $n = 48$ cells; MIA offspring + LPS + anti-IL-17a, $n = 43$ cells; from 2 PBS offspring and 2 MIA offspring in 12 independent experiments. **e**, Lentivirus encoding either EYFP or enhanced green fluorescent protein (EGFP) fused to nuclear Cre (nCre) was bilaterally injected into the S1DZ of *Il-17Ra*$^{fl/fl}$ MIA offspring. Scale bar, 200 µm. **f**, **g**, Mice were tested for sociability one day before injection (pre). The following day, mice were tested for sociability four hours after LPS injection (test). For experiments in **e**, **f**, *Il-17ra*$^{fl/fl}$*;EYFP*, $n = 9$ and *Il-17ra*$^{fl/fl}$*;EGFP:nCre*, $n = 10$; from 5 independent experiments. Unless otherwise indicated, *n* values refer to the number of mice used. Statistics calculated by two-way repeated-measures ANOVA with Bonferroni's post-hoc test (**a**, **f**) or two-way ANOVA with Tukey's post-hoc test (**d**). Graphs are mean ± s.e.m.

on FOS expression in MIA offspring; the number of FOS$^+$ neurons either remained unchanged (as in several of the cortical regions we examined) or increased, as in the central amygdala (CeA)—a region that is known to be activated by LPS[20] (Fig. 2a, c, Extended Data Fig. 6d, e). Therefore, the LPS-induced behavioural rescue in MIA offspring was accompanied by a reduction in neural activity in the S1DZ.

Dysregulation of neural activity and deficits in interneuron function in the S1 have previously been associated with various genetic mouse models of neurodevelopmental disorders[21–23]. We therefore sought to determine whether increased neural activity can also be observed in the S1DZ of monogenic mutant mice. The number of FOS$^+$ neurons in the S1DZ of these mutant mice was increased compared to that of wild-type mice, and the magnitude of this increase correlated with the severity of the sociability deficits—notably in *Cntnap2* and *Fmr1* mutant mice (Extended Data Fig. 7a, b). These data suggest that increased neural activity in the S1DZ may also contribute to the expression of sociability deficits in monogenic mutant mice, similarly to what has been described for MIA offspring[8]. Consistent with this idea, optogenetically reducing neural activity in the S1DZ could rescue sociability deficits in *Cntnap2* and *Fmr1* mutant mice (Fig. 2d–f, Extended Data Fig. 7c–f). *Shank3* mutant mice also showed an increase in sociability upon photoinhibition, but this increase was not significantly different from that of *Shank3* control mice that express enhanced yellow fluorescent protein (EYFP). Therefore, reducing neural activity in the S1DZ was sufficient to restore sociability in *Cntnap2* and *Fmr1* mutant mice, as well as in MIA offspring[8]. Unlike in MIA offspring however, LPS treatment did not reduce the number of FOS$^+$ neurons in the S1DZ of monogenic mutant mice (Fig. 2g–i).

IL-17a rescues sociability deficits

The injection of LPS is known to increase the production of inflammatory cytokines[24]. We observed that administration of LPS results in a robust increase in the plasma levels of IFNγ, IL-6 and TNF (also known as TNF-α) (Fig. 3a). IL-17a—the orthologue of which in *Caenorhabditis elegans* has previously been implicated in modulating sensory responses[25]—was prominently upregulated in MIA offspring, but not in monogenic mutant mice or in control mice (Fig. 3a). Furthermore, we noted that the receptor subunit A for IL-17a (IL-17Ra) is expressed in cortical neurons, including in the S1DZ (Fig. 3b–d, Extended Data Fig. 8a–c). These data suggested that the increased levels of IL-17a after treatment with LPS in MIA offspring, but not in monogenic mutant mice, may directly affect the S1DZ and thus restore sociability. Consistent with this idea, the direct administration of recombinant IL-17a into the S1DZ was sufficient to increase sociability not only in MIA offspring but also in *Cntnap2* and *Fmr1* mutant mice (Fig. 3e, f, Extended Data Fig. 8d–h).

LPS-induced rescue requires IL-17a

To further determine whether IL-17a mediates the LPS-driven behavioural rescue, we inhibited IL-17a activity in the brain via intracerebroventricular injection of blocking antibodies. Antibodies against IL-17a prevented both the LPS-induced rescue of sociability (Fig. 4a, Extended Data Fig. 9a–e) and the reduction of FOS expression in the S1DZ of MIA offspring (Extended Data Fig. 9f). To directly assay the effects of LPS on neural activity, we used multi-electrode arrays to measure the firing rate of neurons of the S1DZ in awake mice

(Extended Data Fig. 10a, b). Upon treatment with LPS, we observed a decrease in the overall firing rate that was prevented by blocking IL-17a in MIA offspring. Treatment with LPS or IL-17a blocking antibody did not change the neuronal firing rate in the S1DZ of control offspring (Fig. 4b–d, Extended Data Fig. 10c). Furthermore, LPS injection did not restore sociability to MIA offspring that are deficient for IL-17Ra in the neurons of the S1DZ (Fig. 4e, f, Extended Data Fig. 10d–j). Our data collectively demonstrate that IL-17a mediates the restoring effects of inflammation on social behaviours by directly acting on IL-17Ra^{+} neurons of the S1DZ.

Discussion

Previous data in mice have suggested that increased production of IL-17a in pregnant mothers may present as a risk factor for neurodevelopmental disorders in offspring[15,26,27]. On the basis of our current findings, we propose that in adult MIA offspring the same cytokine is beneficial and ameliorates sociability phenotypes during episodes of inflammation. Treatment with LPS led to an increase in IL-17a levels in the blood selectively in MIA offspring—but not in other monogenic mutant mice. This suggests that the inflammatory responses may result in beneficial effects only for individuals who have their immune systems primed by prenatal exposure to immune activation, or by other environmental factors. A better understanding of the role of a primed immune system among patients with neurodevelopmental disorders may help to identify those patients whose behavioural symptoms are likely to improve after exposure to fever-associated inflammation. Furthermore, elucidating the mechanisms by which IL-17a can induce two opposing behavioural outcomes depending on when its upregulation occurs (during embryonic brain development or in the adult brain) may provide opportunities to devise therapeutic as well as preventive treatments for the behavioural symptoms associated with ASD.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1843-6.

1. Curran, L. K. et al. Behaviors associated with fever in children with autism spectrum disorders. *Pediatrics* **120**, e1386–e1392 (2007).
2. Grzadzinski, R., Lord, C., Sanders, S. J., Werling, D. & Bal, V. H. Children with autism spectrum disorder who improve with fever: insights from the Simons Simplex Collection. *Autism Res*. **11**, 175–184 (2018).
3. Shi, L., Fatemi, S. H., Sidwell, R. W. & Patterson, P. H. Maternal influenza infection causes marked behavioral and pharmacological changes in the offspring. *J. Neurosci*. **23**, 297–302 (2003).
4. Smith, S. E. P., Li, J., Garbett, K., Mirnics, K. & Patterson, P. H. Maternal immune activation alters fetal brain development through interleukin-6. *J. Neurosci*. **27**, 10695–10702 (2007).
5. Peñagarikano, O. et al. Absence of CNTNAP2 leads to epilepsy, neuronal migration abnormalities, and core autism-related deficits. *Cell* **147**, 235–246 (2011).
6. The Dutch-Belgian Fragile X Consortium. *Fmr1* knockout mice: a model to study fragile X mental retardation. *Cell* **78**, 23–33 (1994).
7. Peça, J. et al. *Shank3* mutant mice display autistic-like behaviours and striatal dysfunction. *Nature* **472**, 437–442 (2011).
8. Yim, Y. et al. Reversing behavioural abnormalities in mice exposed to maternal inflammation. *Nature* **549**, 482–487 (2017).
9. Whitrow, M. Wagner-Jauregg and fever therapy. *Med. Hist*. **34**, 294–310 (1990).
10. Lyall, K. et al. The changing epidemiology of autism spectrum disorders. *Annu. Rev. Public Health* **38**, 81–102 (2017).
11. Crawley, J. N. Translational animal models of autism and neurodevelopmental disorders. *Dialogues Clin. Neurosci*. **14**, 293–305 (2012).
12. Ey, E., Leblond, C. S. & Bourgeron, T. Behavioral profiles of mouse models for autism spectrum disorders. *Autism Res*. **4**, 5–16 (2011).
13. Patterson, P. H. Maternal infection and immune involvement in autism. *Trends Mol. Med*. **17**, 389–394 (2011).
14. Kozak, W., Conn, C. A. & Kluger, M. J. Lipopolysaccharide induces fever and depresses locomotor activity in unrestrained mice. *Am. J. Physiol*. **266**, R125–R135 (1994).
15. Choi, G. B. et al. The maternal interleukin-17a pathway in mice promotes autism-like phenotypes in offspring. *Science* **351**, 933–939 (2016).
16. Malkova, N. V., Yu, C. Z., Hsiao, E. Y., Moore, M. J. & Patterson, P. H. Maternal immune activation yields offspring displaying mouse versions of the three core symptoms of autism. *Brain Behav. Immun*. **26**, 607–616 (2012).
17. Armbruster, B. N., Li, X., Pausch, M. H., Herlitze, S. & Roth, B. L. Evolving the lock to fit the key to create a family of G protein-coupled receptors potently activated by an inert ligand. *Proc. Natl Acad. Sci. USA* **104**, 5163–5168 (2007).
18. Vong, L. et al. Leptin action on GABAergic neurons prevents obesity and reduces inhibitory tone to POMC neurons. *Neuron* **71**, 142–154 (2011).
19. Zhao, Z. D. et al. A hypothalamic circuit that controls body temperature. *Proc. Natl Acad. Sci. USA* **114**, 2042–2047 (2017).
20. Cai, H., Haubensak, W., Anthony, T. E. & Anderson, D. J. Central amygdala PKC-δ^{+} neurons mediate the influence of multiple anorexigenic signals. *Nat. Neurosci*. **17**, 1240–1248 (2014).
21. Gogolla, N. et al. Common circuit defect of excitatory-inhibitory balance in mouse models of autism. *J. Neurodev. Disord*. **1**, 172–181 (2009).
22. Selby, L., Zhang, C. & Sun, Q. Q. Major defects in neocortical GABAergic inhibitory circuits in mice lacking the fragile X mental retardation protein. *Neurosci. Lett*. **412**, 227–232 (2007).
23. Orefice, L. L. et al. Targeting peripheral somatosensory neurons to improve tactile-related phenotypes in ASD models. *Cell* **178**, 867–886 (2019).
24. Erickson, M. A. & Banks, W. A. Cytokine and chemokine responses in serum and brain after single and repeated injections of lipopolysaccharide: multiplex quantification with path analysis. *Brain Behav. Immun*. **25**, 1637–1648 (2011).
25. Chen, C. et al. IL-17 is a neuromodulator of *Caenorhabditis elegans* sensory responses. *Nature* **542**, 43–48 (2017).
26. Kim, S. et al. Maternal gut bacteria promote neurodevelopmental abnormalities in mouse offspring. *Nature* **549**, 528–532 (2017).
27. Lammert, C. R. et al. Cutting edge: critical roles for microbiota-mediated regulation of the immune system in a prenatal immune activation model of autism. *J. Immunol*. **201**, 845–850 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Mice

All experiments were performed according to the Guide for the Care and Use of Laboratory Animals and were approved by the National Institutes of Health and the Committee on Animal Care at Massachusetts Institute of Technology. C57BL/6 mice were purchased from Taconic Biosciences for generating PBS and MIA offspring. *Slc32a1*(*Vgat*)–Cre (028862) mice were purchased from Jackson Laboratories, and inbred. For monogenic experiments, C57BL/6, *Cntnap2* (017482), *Fmr1* (003025) and *Shank3* (017688) mice were purchased from Jackson Laboratories, and inbred after colonizing with segmented filamentous bacteria (SFB) from donor mice. *Il-17ra*$^{fl/fl}$ and IL-17Ra-knockout mice have previously been described[28,29]. All mice were males aged 2–5 months, unless otherwise specified.

Maternal immune activation

Mice were mated overnight with females carrying SFB in their guts[26]. On E12.5, pregnant female mice were weighed and injected with a single dose (20 mg/kg, intraperitoneal injection) of poly(I:C) (P9582, Sigma Aldrich) or PBS. Each dam was returned to its cage and left undisturbed until the birth of its litter. All pups remained with the mother until weaning on postnatal day (P)21–28, at which time mice were group-housed at a maximum of 5 per cage with same-sex littermates. Matings between *Vgat*–Cre (c/c) males and wild-type females were used to make MIA *Vgat*–Cre mice.

Stereotaxic surgery

Surgeries were carried out using aseptic techniques. Mice were anaesthetized using a mixture of ketamine (100 mg/kg, intraperitoneal injection) and xylazine (10 mg/kg, intraperitoneal injection). Mice were given pre-operative slow-release buprenorphine (1.0 mg/kg, subcutaneous injection). For manipulating body temperature with vLPO inhibition, *Vgat*–Cre mice received bilateral stereotaxic injections of virus (200 nl per side) at rates of <0.1 ml/min (AAV_2-hSyn-DIO-hM4Di-mCherry (Addgene)). Virus was bilaterally targeted to the vLPO (AP +0.15, medial–lateral (ML) ±0.50, dorsal–ventral (DV), −4.9, lambda was raised 600 µm above bregma). For experiments involving optical manipulation in the S1DZ, wild-type, *Cntnap2*, *Fmr1* and *Shank3* mice received bilateral stereotaxic injections of virus (400 µl/side) at a rate of <0.1 ml/min (AAV_2-hSyn-EYFP, AAV_2-hSyn-eNpHR3.0:EYFP). Viruses were bilaterally targeted to the S1DZ (AP −0.50, ML ±2.6, DV −0.80, lambda was level with bregma). To deliver light into the S1DZ, a 300-µm optic fibre was superficially implanted in the S1DZ. For experiments involving genetic knockout of IL-17Ra in the S1DZ, *Il-17ra*$^{fl/fl}$ mice received bilateral stereotaxic injections of virus (800 nl per side) at a rate of <0.1 ml/min (pLenti-hSyn-EYFP, pLenti-hSyn-EGFP:nCre). Viruses were bilaterally targeted to the S1DZ. All lentiviruses were made in-house. For administration of cytokines into the S1DZ, cannulas (PlasticsOne) were implanted superficially within the S1DZ. For central administration of blocking antibodies, cannulas were implanted above the right lateral ventricle (AP −0.30, ML +1.0, DV −1.35, lambda was raised 600 µm above bregma). Cannulas were fitted with dummy cannulas (PlasticsOne) to maintain cannula patency after surgery.

Tracking of body temperature

Body temperature was measured using the Anipill remote temperature monitoring system (007894-001, DSI). Adult male mice were implanted with an Anipill capsule in the abdominal cavity. Experiments were carried out more than three weeks after surgery. Mice were singly housed the day before the experiment. Ambient temperature was maintained at 23.5 °C, consistent with the ambient temperature of the vivarium. Body temperature was sampled in 5-min increments. LPS or vehicle injections occurred between 11:00–13:00 for experiments assaying the effects of LPS on body temperature in PBS and MIA offspring.

Immunohistochemistry

Mice were transcardially perfused with cold paraformaldehyde (PFA) (4% in PBS). Brains were kept in PFA overnight at 4 °C before vibratome-sectioning (Leica VT1000s). Brains were cut at 50-µm thickness for FOS quantification. Brains were cut at 100-µm thickness for all other experiments.

Before antibody labelling, sections were incubated in blocking solution (0.4% Triton X-100 and 2% goat serum in PBS) for 30 min. Sections were then incubated in blocking solution containing primary antibodies overnight at room temperature. Primary antibodies used were chicken anti-GFP (1:1,000, Ab5450, Abcam), rabbit anti- FOS (1:500, ABE457, Millipore), rabbit anti-DsRed (1:1,000, 632496, Clontech) and mouse anti-NeuN (1:1,000, MAB377, Millipore). Sections were washed in wash buffer (0.4% Triton X-100 in PBS) three times before secondary antibody labelling. Sections were incubated in blocking solution containing secondary antibodies and DAPI (1:5,000, D1306, Thermo Fisher) for three hours at room temperature. Images of stained slices were acquired using a confocal microscope (LSM710, Carl Zeiss) with a 10×, 20× or 40× objective lens.

In situ hybridization

Mice were transcardially perfused with cold PBS. Brains were extracted and embedded in optimal cutting temperature (OCT) compound on dry ice. Sections were cut at 20-µm thickness on a cryostat. In situ hybridizations were performed using the RNAscope 2.5 HD Assay-Red kit (322350, Advanced Cell Diagnostics) using a probe targeting the *Il17ra* transcript (Mm-Il17ra-O1, 566131, Advanced Cell Diagnostics). The probe was designed to target region 444–882 of the *Il17ra* transcript (NM_008359.2). Modifications to the kit protocol to improve adherence of tissue to the slide included an extension of the fixation time to 30 min and the addition of a humidified bake step at 40 °C immediately before probe hybridization. Sections were counterstained with DAPI. Images were acquired using a confocal microscope (LSM710, Carl Zeiss) with a 10× or 20× objective lens. *Il17ra* and DAPI expression was quantified using QuPath[30]. Cells were divided into the following categories based on level of *Il17ra* expression: low = 1–3 puncta, medium = 4–9 puncta, high = 10–15 puncta, highest = >15 puncta.

In situ hybridization followed by immunohistochemistry

For experiments assaying the overlap of *Il17ra* and NeuN expression, immunohistochemistry for NeuN was performed following a modified in situ hybridization protocol. The RNAscope 2.5 HD Assay-Red kit in situ protocol was modified in the following ways: sections were baked at 60 °C, followed by a 10-min fixation step with 4% PFA at room temperature. Sections were then stored in 70% ethanol at 4 °C overnight. Sections were permeabilized in 8% SDS for 10 min. Sections were washed twice with PBS between each step. After SDS treatment, the RNAscope 2.5 HD Assay-Red kit protocol was followed from the probe hybridization step. Following the completion of the in situ protocol, sections were incubated in blocking buffer containing anti-NeuN antibody overnight at 4 °C. Sections were then incubated in blocking buffer containing DAPI and secondary antibody for two hours at room temperature.

Behavioural analysis

Male mice were tested during the light cycle in a room with lighting maintained at 230 lux. Mice were transferred to the testing area at least one hour before the initiation of experiments. Tracking of mouse behaviour was done using the EthoVision XT (Noldus) tracking system.

Three-chamber social approach assay. Adult male mice were assayed for sociability using a three-chamber social approach assay. The arena

was constructed of white acrylic (50 cm × 35 cm × 30 cm). Wire cups (Spectrum Diversified) were placed in the back left and right corner of the arena beneath water-filled 1-l bottles (Nalgene). On day 0, mice were habituated to the arena for 10 min. Immediately after habituation, mice were singly housed. On day 1 (pre-test), mice were placed in the centre of the arena and allowed to freely explore. Following 10 min, mice were confined to the centre of the arena. An inanimate object (rubber stopper) or a male conspecific were placed beneath the wire cups. Placement of the inanimate object and social target were alternated. Mice were then allowed to freely explore the arena for 10 min. Interaction time was defined as time spent in the areas circumscribing the wire cups (<2 cm). Sociability was defined as interaction time with the social target divided by total interaction time and expressed as a percentage. For experiments involving LPS injections, mice were injected with either saline (vehicle) or LPS (50 µg/kg, intraperitoneal injection, L2630, Sigma) on day 2 (test), four hours before testing. For the experiment assaying sociability 72 h after injection of LPS, mice used for 4-h LPS sociability experiments were tested for sociability again at 72 h.

Three-chamber social approach assay with DREADD manipulation. Adult male *Vgat*–Cre MIA offspring were bilaterally injected with virus encoding the inhibitory DREADD receptor fused to mCherry into the vLPO. After more than three weeks of recovery, mice were assayed on the three-chamber social approach assay outlined above. Baseline sociability was assayed on day 1 (pre-test). On day 2 and day 3, mice were injected with either vehicle or CNO (1.5 mg/kg, intraperitoneal injection, BML-NS105, Enzo Life Sciences), two hours before initiation of behaviour. Injection order was counterbalanced. Following behavioural experiments, post-mortem histology was used to confirm mCherry expression within the vLPO. For experiments assaying the effect of injection of CNO and LPS in *Vgat*–Cre PBS and MIA offspring that have not undergone surgery, baseline sociability was assessed on day 1. On day 2 and 3, mice received counterbalanced injections of CNO or vehicle. On day 4, mice were injected with LPS.

Three-chamber social approach assay with administration of IL-17a into the S1DZ. Adult male mice were implanted with a cannula into S1DZ bilaterally and allowed to recover for more than two weeks before the behavioural experiments. On day 1 (pre-test), mice were assayed for sociability. On day 2 (test), mice were anaesthetized briefly using isoflurane and either vehicle or IL-17a (50 ng per side in 1 µl at a rate of 180 nl/min, 7956-ML/CF, R&D) was administered bilaterally into the S1DZ through 250-µm projecting injector tips (PlasticsOne). Four hours after administration of vehicle or IL-17a, mice were assayed for sociability. Cannula placements were verified using histology.

Three-chamber social approach assay with S1DZ optogenetic inhibition. EYFP or NpHR were virally targeted to the S1DZ. After two weeks of recovery, mice were allowed to freely explore the arena for 10 min. The following day, the mice were given 3 min of no stimulation ('off' session) and 3 min of laser stimulation ('on' session) (594 nm, 6 mW).

Three-chamber social approach assay with IL-17a blocking antibody. For central cytokine blockade experiments, adult male mice were implanted with a cannula into the lateral ventricle and allowed to recover for more than two weeks before the behavioural experiments. On day 1 (pre-test), mice were tested for baseline sociability. On day 2 (test), mice were injected with IL-17a blocking antibody (clone 50104; R&D) or isotype-control antibody (IgG2a, clone 54447; R&D). Antibodies were dissolved in saline and administered intracerebroventricularly at 1 mg/kg in 500 nl at a rate of 180 nl/min through 750-µm projecting injector tips (PlasticsOne). Blocking antibody was administered 30 min before administration of LPS. Four hours following administration of LPS, mice were assayed for sociability.

Three-chamber social approach assay in mice deficient for IL-17Ra in the S1DZ. IL-17Ra-knockout was mediated by viral delivery of Cre recombinase, expressed under the control of the human synapsin (hSyn) promoter, into the S1DZ of *Il-17ra*$^{fl/fl}$ MIA offspring. After at least three weeks of recovery, mice were assayed for sociability. On the next day sociability was assessed four hours following administration of LPS.

Marble burying assay. On day 1, mice were tested for their baseline marble burying phenotype. On day 2, four hours before beginning the marble burying assay, mice were treated with either LPS or vehicle. The marble burying assay was carried out as previously described[8]. Mice were placed into testing arenas (arena size 40 cm × 20 cm × 30 cm; bedding depth 3 cm) each containing 20 glass marbles (laid out in 4 rows of 5 marbles equidistant from one another). At the end of the 15-min exploration period mice were carefully removed from the testing cages and the number of marbles buried was recorded. The marble burying index was arbitrarily defined as the following: 1 for marbles covered >50% with bedding, 0.5 for marbles covered <50% with bedding or 0 for anything less.

Reciprocal social interaction assay. Four hours before testing, mice were injected with vehicle or LPS. Two unfamiliar mice of the same treatment and background were placed in a fresh mouse cage and allowed to freely interact for 10 min. Videos were acquired using IC Capture (The Imaging Source) at a 640 × 480 aspect ratio and 25 frames per second. Social interaction (close following, push–crawl, nose–nose sniffing and nose–anus sniffing) was scored by an observer blind to treatment and background[31].

Quantification of FOS$^+$ cells in the brain after administration of LPS

Adult male mice were killed five hours after injection of LPS. FOS$^+$ cells were quantified using the Cell Counter plugin in Fiji[32]. All cells were counted within a single coronal section of each respective brain region, as defined by the Paxinos and Franklin Mouse Atlas[33]. Regions quantified include: medial prefrontal cortex (mPFC) (comprising prelimbic cortex (PrL) and and infralimbic cortext (IL)), AP +1.98; S1DZ, AP −0.46; primary somatosensory cortex barrel field (S1BF), AP −0.46; primary motor cortex (M1), AP −0.46; secondary motor cortex (M2), AP −0.46; secondary auditory cortex dorsal part (AuD), AP −1.94; CeA, AP −1.94; and primary visual cortex (V1), AP −3.64. For experiments testing the IL-17a dependence of LPS-induced changes in FOS expression, mice were injected intracerebroventricularly with control antibodies or blocking antibody against IL-17a 30 min before intraperitoneal injection of vehicle or LPS. Surgical and injection methods were identical to behavioural experiments.

Enzyme-linked immunosorbent assay

After four hours of administration of vehicle or LPS, mice were anaesthetized by intraperitoneal injection of Fatal-Plus (100 mg/kg). All blood samples were centrifuged at 10,000*g* for 10 min at 4 °C. All samples were stored at −80 °C until further analysis. Cytokine concentrations in plasma were measured using an ELISA kit (IFNγ; 430804, TNF; 430904, IL-6; 431304 and IL-17a; 432504, Biolegend), following the manufacturer's instructions.

PCR for assaying IL-17Ra knockout

Il17ra$^{fl/fl}$ male mice were bilaterally injected with virus encoding nuclear Cre fused to EGFP, or control virus encoding only EYFP, into the S1DZ. After more than three weeks, injection sites were dissected from the S1. Single cells were dissociated from brain tissue using a modified version[34] of the Papain Dissociation Kit protocol (LK003153, Worthington) and sorted on a BD FACS Aria (BD Biosciences) based on EGFP and EYFP expression. RNA was extracted from sorted cells using

a Quick-RNA micro-prep kit (Zymo). Twenty nanograms of RNA was converted into cDNA using oligodT (Protoscript First Strand CDNA Synthesis Kit, NEB). *Il17ra* mRNA expression was analysed using PCR. One microlitre of cDNA was diluted in a 20-µl reaction volume. *Il17ra* and *Gapdh* mRNA expression was assessed using the following primers: *Il17ra* 5′-AGATGCCAGCATCCTGTACC-3′ and 5′-CACAGTCACAG CGTGTCTCA-3′; *Gapdh* 5′-GACTTCAACAGCCTCCCACTCTTCC-3′ and 5′-TGGGTGGTCCAGGTTTCTTACTCCTT-3′. Cycling conditions for *Il17ra*: 95 °C × 5 min (1 cycle), 95 °C × 20 s, 60 °C × 30 s, 72 °C × 30 s (32 cycles), 72 °C × 5 min (1 cycle) and 4 °C hold. Cycling conditions for *Gapdh*: 95 °C × 5 min (1 cycle), 95 °C × 20 s, 60 °C × 30 s, 72 °C × 30 s (28 cycles), 72 °C × 5 min (1 cycle) and 4 °C hold. Band intensity from gel images was quantified using ImageJ.

In vivo electrophysiology

Electrophysiological experiments were conducted in head-fixed mice trained to walk on a rotating running wheel. Before training, mice were implanted with custom crowns permitting head-fixing above the wheel and allowed to recover for one week before training. Mice were trained to walk on the wheel for three 10-min sessions daily for at least one week. Following training, mice were implanted with a multi-electrode array targeting the S1DZ and allowed to recover before testing. To assess the IL-17a dependence of LPS-induced changes in neural activity, mice were injected intraperitoneally with control antibodies or blocking antibody against IL-17a (1 mg per mouse), 30 min before intraperitoneal injection of vehicle or LPS. Baseline neural activity was measured while mice were running on the wheel immediately before the first injection. After the first injection, mice were returned to their home cage. Post-injection (test) neural activity during wheel running was measured 4 h after injection of vehicle or LPS.

Multi-electrode array construction and implantation. Custom multi-electrode array scaffolds (drive bodies) were designed using 3D CAD software (SolidWorks) and printed in Accura 55 plastic (American Precision Prototyping), as previously described[35,36]. Before implantation, each array scaffold was loaded with 8–24 independently movable microdrives carrying 12.5-µm nichrome (California Fine Wire Company) tetrodes. Electrodes were pinned to custom-designed 32- or 128-channel electrode interface boards (Sunstone Circuits) along with a common reference wire (A-M systems).

Electrophysiological recordings and spike sorting. Signals were acquired using a Neuralynx multiplexing digital recording system (Neuralynx) through a combination of 32- and 64-channel digital multiplexing headstages plugged into the electrode interface board of the implant. Signals from each electrode were amplified, filtered between 0.1 Hz and 9 kHz and digitized at 30 kHz. Initial spike sorting was performed using MountainSort, followed by manual quality control using MClust toolbox (http://redishlab.neuroscience.umn.edu/mclust/MClust.html).

Analysis of firing rate. At time 0 in each recording session, mice began walking on the wheel rotating at 7.5 cm s^{-1}. Data collected between 1 and 2 s were included in the analysis. The firing rate was calculated across 1-s time windows during which mice were walking stably on the rotating wheel, averaged across 25–50 trials per condition. The first second after the onset of wheel rotation was omitted, to avoid firing rate changes due to acceleration. Firing rate was sampled with a 1-ms bin width passed through a box car filter (100 ms). The resulting peristimulus time histograms were then smoothed with a 50-ms Gaussian. To assess the effect of treatment on neural activity, changes in firing rates during wheel running 4 h after injection were normalized to the pre-injection baseline. Cells showing a firing rate below 0.1 Hz were excluded from analysis.

Statistics and reproducibility

Statistical analyses were performed using GraphPad Prism. Sample size was chosen on the basis of similar previous studies, and not on statistical methods to predetermine sample size[8,15]. Within each iteration of an experiment, mice were randomly assigned to groups with approximately balanced sample size. Behavioural results from mice with inaccurate targeting of viral infection or cannula implantations were excluded. Experimenters were blind to subject treatment during data collection and analysis.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Source Data for Figs. 1–4 and Extended Data Figs. 1–4, 6–10, containing raw data for all experiments, are provided with the paper. All other data are available from the corresponding author on reasonable request.

28. El Malki, K. et al. An alternative pathway of imiquimod-induced psoriasis-like skin inflammation in the absence of interleukin-17 receptor a signaling. *J. Invest. Dermatol.* **133**, 441–451 (2013).
29. Tusi, B. K. et al. Population snapshots predict early haematopoietic and erythroid hierarchies. *Nature* **555**, 54–60 (2018).
30. Bankhead, P. et al. QuPath: open source software for digital pathology image analysis. *Sci. Rep.* **7**, 16878 (2017).
31. Silverman, J. L., Yang, M., Lord, C. & Crawley, J. N. Behavioural phenotyping assays for mouse models of autism. *Nat. Rev. Neurosci.* **11**, 490–502 (2010).
32. Schindelin, J. et al. Fiji: an open-source platform for biological-image analysis. *Nat. Methods* **9**, 676–682 (2012).
33. Paxinos, G. & Franklin, K. B. J. *The Mouse Brain in Stereotaxic Coordinates* (Academic, 2001).
34. Hrvatin, S. et al. Single-cell analysis of experience-dependent transcriptomic states in the mouse visual cortex. *Nat. Neurosci.* **21**, 120–129 (2018).
35. Halassa, M. M. et al. State-dependent architecture of thalamic reticular subnetworks. *Cell* **158**, 808–821 (2014).
36. Brunetti, P. M. et al. Design and fabrication of ultralight weight, adjustable multi-electrode probes for electrophysiological recordings in mice. *J. Vis. Exp.* **91**, e51675 (2014).

Acknowledgements We thank N. Soares, M. Garcia and Y. Liu for assistance with experiments, and B. Noro and M. Trombly for critical reading of the manuscript. This work was supported by the National Institute of Mental Health (1-R01-MH115037-01, to G.B.C.), Jeongho Kim Neurodevelopmental Research Fund (G.B.C. and J.R.H.), Hock E. Tan and K. Lisa Yang Center for Autism Research (G.B.C. and M.D.R.), P. Ha (G.B.C.), Simons Center for the Social Brain (G.B.C., J.R.H. and Y.S.Y.), the Simons Foundation Autism Research Initiative (G.B.C. and J.R.H.), the Champions of the Brain Weedon Fellowship (G.M.W.) and the National Science Foundation Graduate Research Fellowship (no. 1122374, to M.D.R.). We are deeply grateful to B. Picower, the JPB Foundation, the Picower Institute for Learning and Memory, Lore McGovern and the McGovern Institute for Brain Research for their mentorship and direct support of this work over the years.

Author contributions M.D.R., Y.S.Y., M.M.H., J.R.H. and G.B.C. designed the experiments and/or provided advice and technical expertise. M.D.R., Y.S.Y., R.D.W., H.K., C.R., G.M.W., M.A. and H.O.K. performed the experiments. A.W. provided *Il-17ra*$^{fl/fl}$ mice. M.D.R., Y.S.Y., J.R.H. and G.B.C. wrote the manuscript with input from the co-authors.

Competing interests The authors declare no competing interests.

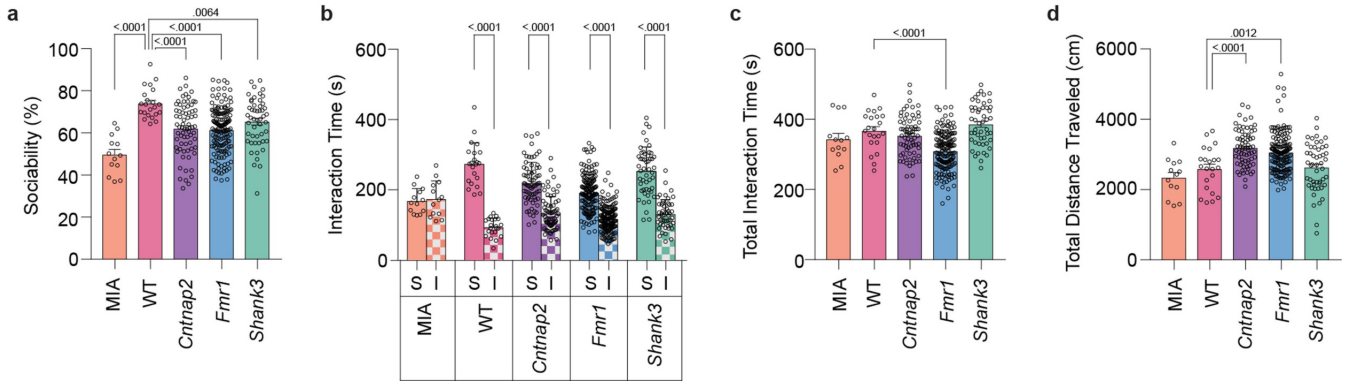
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1843-6.

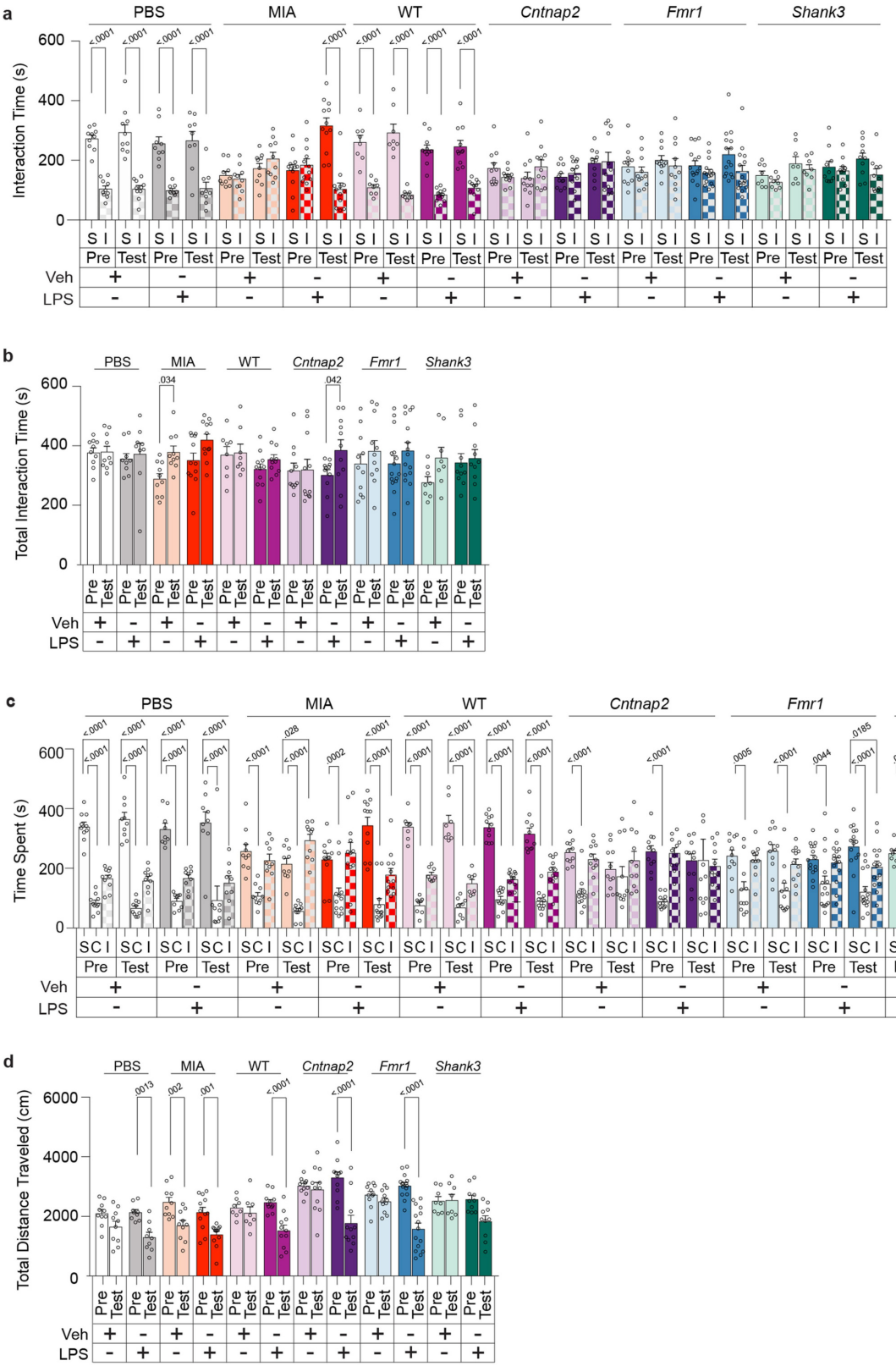
Correspondence and requests for materials should be addressed to J.R.H. or G.B.C.

Peer review information *Nature* thanks Thomas Blank, Craig M. Powell, Marco Prinz and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

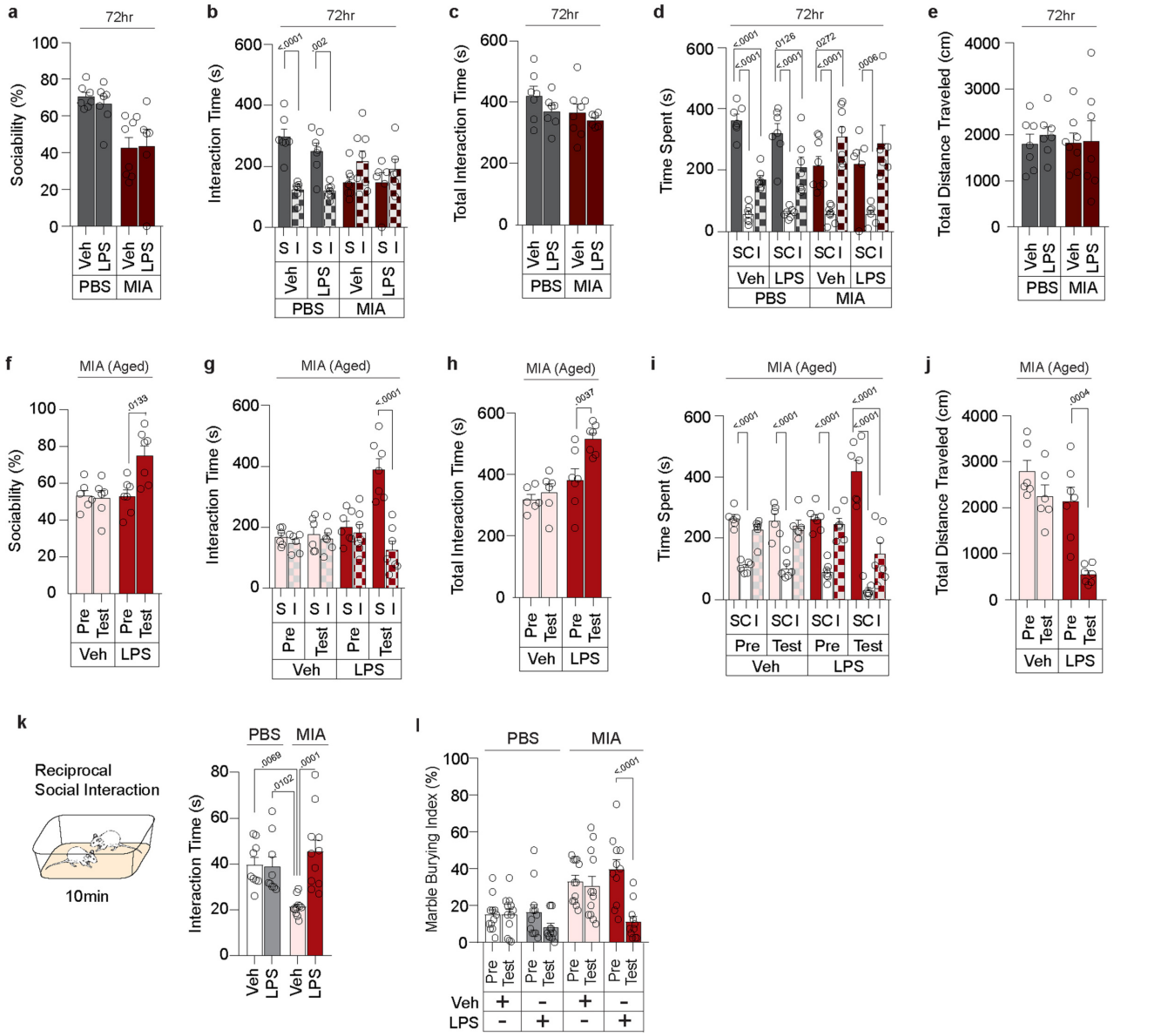
Reprints and permissions information is available at http://www.nature.com/reprints.



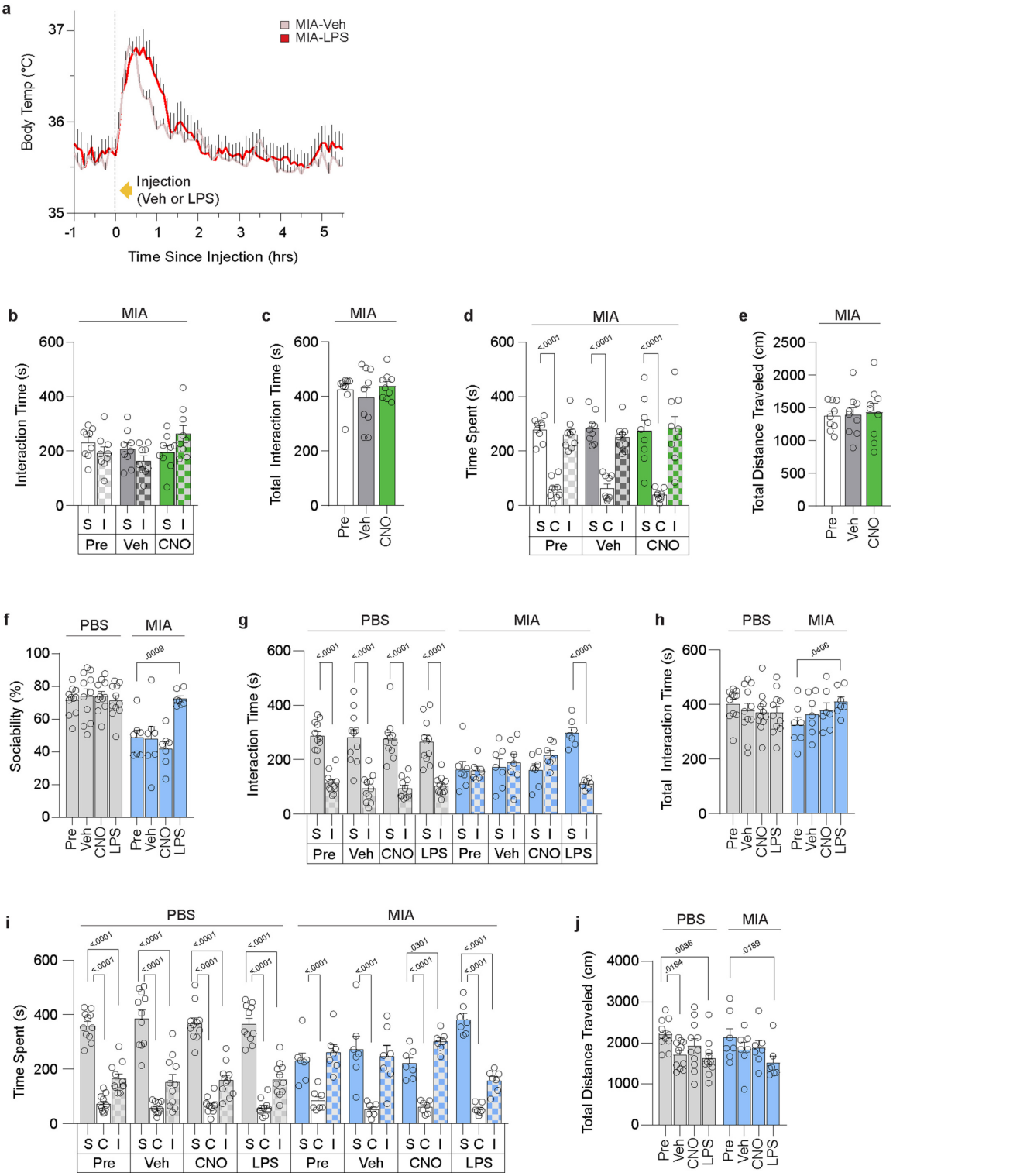
Extended Data Fig. 1 | *Cntnap2*, *Fmr1* and *Shank3* mutant mice show variable sociability performance. a, Sociability performance. MIA, $n = 13$; wild-type, $n = 22$; *Cntnap2* mutant, $n = 71$; *Fmr1* mutant, $n = 165$; *Shank3* mutant, $n = 50$; from 30 independent experiments. **b**, Time spent investigating social (S) versus inanimate (I) objects for mice described in **a**. **c**, **d**, Total interaction time (**c**) and distance travelled (**d**) during the three-chambered sociability experiments described in **a**. All *n* values refer to the number of mice used. Statistics calculated by one-way ANOVA with Dunnett's post-hoc test (**a**, **c**, **d**) or two-way ANOVA with Dunnett's post-hoc test (**b**). Graphs are mean ± s.e.m.



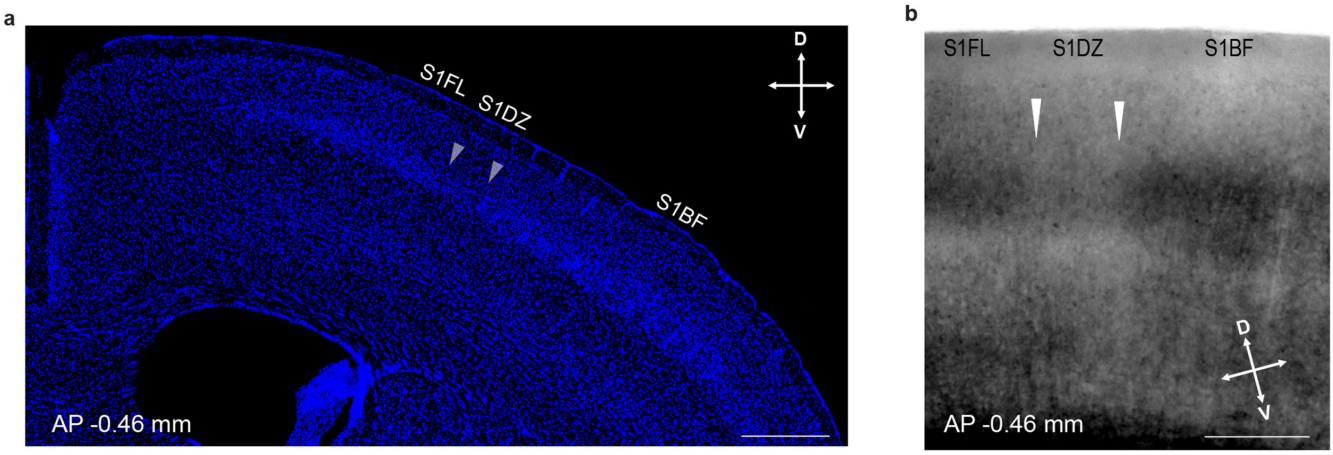
Extended Data Fig. 2 | Further behavioural analyses for sociability performance after treatment with LPS in PBS and MIA offspring, and monogenic mutant mice. a–d, Time spent investigating social (S) versus inanimate (I) objects (**a**), total interaction time (**b**), time spent in social (S), centre (C) or inanimate (I) chamber (**c**), and distance travelled (**d**) for sociability experiments in Fig. 1c. PBS offspring + vehicle, $n = 10$; PBS offspring + LPS, $n = 9$; MIA offspring + vehicle, $n = 10$; MIA offspring + LPS, $n = 12$; wild type (WT) + vehicle, $n = 8$; wild type + LPS, $n = 11$; *Cntnap2* mutant + vehicle, $n = 11$; *Cntnap2* mutant + LPS, $n = 11$; *Fmr1* mutant + vehicle, $n = 11$; *Fmr1* mutant + LPS, $n = 15$; *Shank3* mutant + vehicle, $n = 8$; *Shank3* mutant + LPS, $n = 10$; from 3 independent experiments. All n values refer to the number of mice used. Statistics calculated by two-way ANOVA with Sidak's (**a**) or Dunnett's (**c**) post-hoc test, or two-way repeated-measures ANOVA with Sidak's post-hoc test (**b**, **d**). Graphs are mean ± s.e.m.



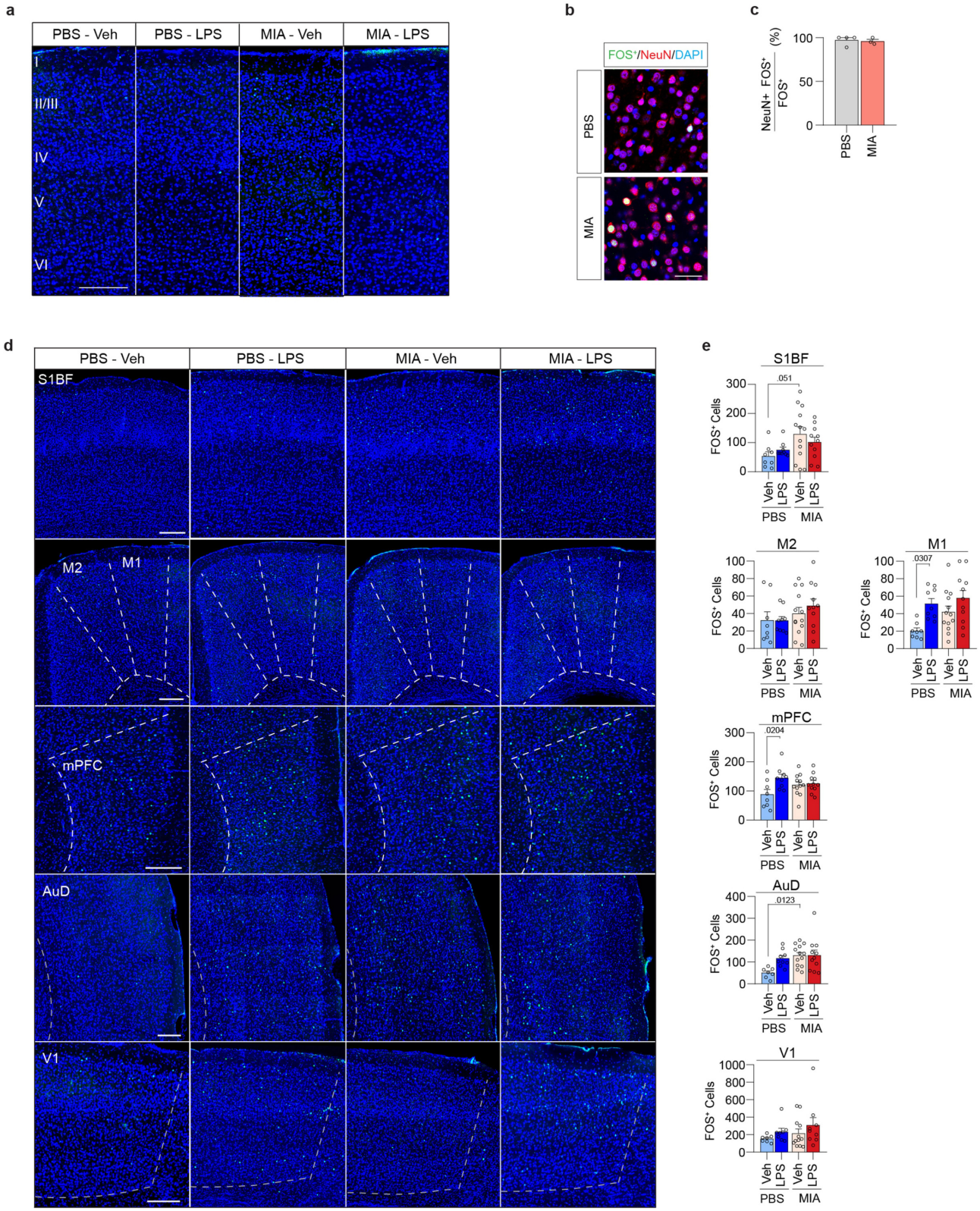
Extended Data Fig. 3 | LPS-induced rescue of MIA behavioural phenotypes is transient, effective in aged mice and extends beyond three-chambered sociability. a–**e**, Sociability measured 72 h after injection with vehicle (Veh) or LPS, in PBS and MIA offspring from Fig. 1c. Data expressed as per cent sociability (**a**), time spent investigating social (S) versus inanimate (I) objects (**b**), total interaction time (**c**), time spent in social (S), centre (C) or inanimate (I) chamber (**d**), and distance travelled (**e**) during three-chambered sociability experiments. PBS offspring + vehicle, $n = 7$; PBS offspring + LPS, $n = 7$; MIA offspring + vehicle, $n = 8$; MIA offspring + LPS, $n = 6$; from 2 independent experiments. **f**–**j**, Sociability measured before and 4 h after injection of vehicle or LPS in aged MIA mice (9–12 months old). Data are expressed as per cent sociability (**f**), time spent investigating social (S) versus inanimate (I) objects (**g**), total interaction time (**h**), time spent in social (S), centre (C) or inanimate (I) chamber (**i**), and distance travelled (**j**) during three-chambered sociability experiments. MIA offspring + vehicle, $n = 6$; MIA offspring + LPS, $n = 7$; from 2 independent experiments. **k**, Reciprocal social interactions measured after treatment with vehicle or LPS in PBS or MIA offspring. PBS offspring + vehicle, $n = 9$; PBS offspring + LPS, $n = 9$; MIA offspring + vehicle, $n = 11$; MIA offspring + LPS, $n = 11$; from 4 independent experiments. **l**, Marble burying index (percentage of buried marbles) measured before and 4 h after treatment with vehicle or LPS in PBS or MIA offspring. PBS offspring + vehicle, $n = 12$; PBS offspring + LPS, $n = 12$; MIA offspring + vehicle, $n = 12$; MIA offspring + LPS, $n = 11$; from 5 independent experiments. All n values refer to the number of mice used. Statistics calculated by two-way ANOVA with Sidak's (**a**–**c**, **e**, **g**), Dunnett's (**d**, **i**) or Tukey's (**k**) post-hoc tests, or two-way repeated-measures ANOVA with Sidak's post-hoc test (**f**, **h**, **j**, **l**). Graphs are mean ± s.e.m.



Extended Data Fig. 4 | Acute increase in body temperature is insufficient to promote sociability. a, Body temperature profile after injection of vehicle or LPS in MIA offspring. Vehicle, $n = 10$; LPS, $n = 10$; from 4 independent experiments. The initial spike in body temperature is due to handling stress. **b**–**e**, Data are expressed as time spent investigating social (S) versus inanimate (I) objects (**b**), total interaction time (**c**), time spent in social (S), centre (C) or inanimate (I) chamber (**d**), and distance travelled (**e**) during three-chambered sociability experiments described in Fig. 1g. $n = 9$ for all groups, from 2 independent experiments. **f**–**j**, Sociability performance in *Vgat*–Cre PBS and MIA offspring after treatment with vehicle, CNO or LPS. Data are expressed as per cent sociability (**f**), time spent investigating social (S) versus inanimate (I) objects (**g**), total interaction time (**h**), time spent in social (S), centre (C) or inanimate (I) chamber (**i**), and distance travelled (**j**) during three-chambered sociability experiments. PBS offspring, $n = 11$; MIA offspring, $n = 7$; from 2 independent experiments. All n values refer to the number of mice used. Statistics calculated by two-way repeated-measures ANOVA with Bonferroni's (**a**) or Dunnett's (**f**, **h**, **j**) post-hoc tests, two-way ANOVA with Sidak's (**b**, **g**) or Dunnett's (**d**, **i**) post-hoc tests, or one-way repeated-measures ANOVA with Tukey's post-hoc test (**c**, **e**). Graphs are mean ± s.e.m.

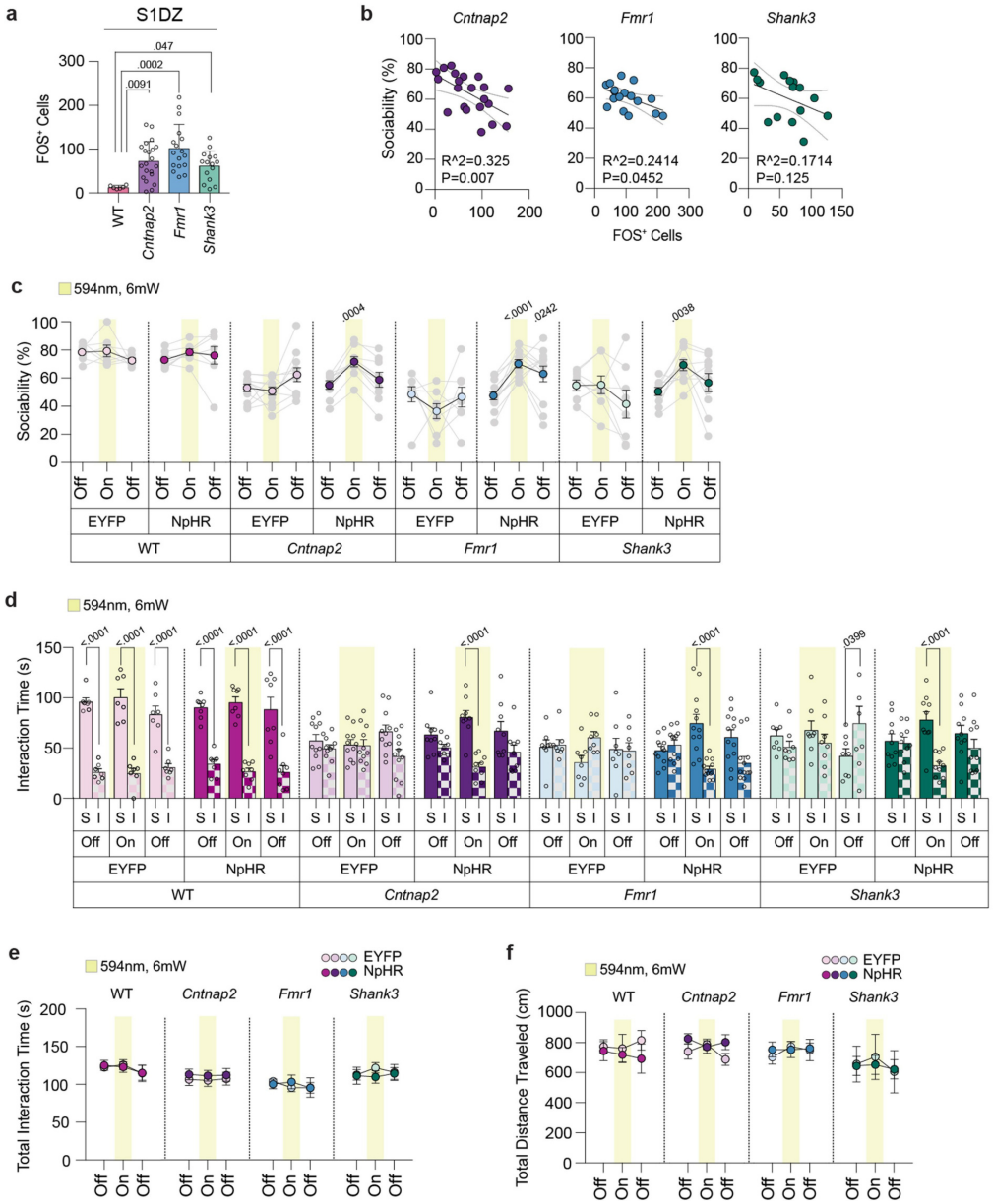


Extended Data Fig. 5 | Histological identification of the S1DZ. a, Coronal section of the cortex counterstained with DAPI to highlight the abrupt reduction in cell density in layer 4, between the S1DZ and the S1BF, at AP −0.46 mm. *n* = 5, from 1 independent experiment. D, dorsal; V, ventral. **b**, Coronal section of the cortex imaged with differential interference contrast, further highlighting the reduced layer 4 in the S1DZ at AP −0.46 mm. *n* = 3, from 2 independent experiments. All *n* values refer to the number of mice used. White arrows indicate borders of S1DZ. Scale bars, 500 µm (**a**); 300 µm (**b**).

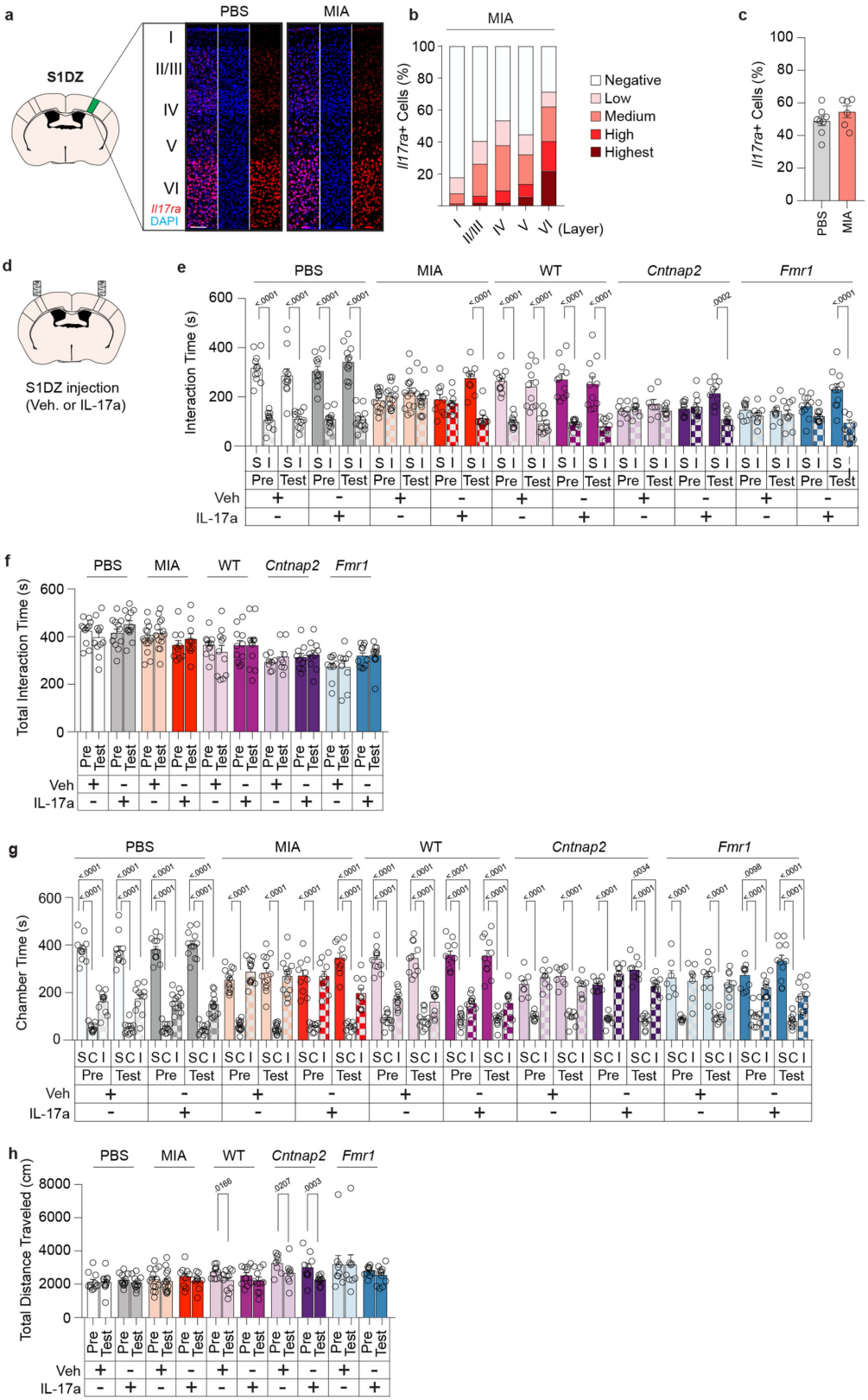


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Treatment of MIA offspring with LPS does not have a distinguishable effect on FOS expression in the other cortical regions analysed. a, Full cortical depth of S1DZ FOS staining as shown in Fig. 2a, for PBS and MIA offspring after administration of vehicle or LPS. PBS offspring + vehicle, $n = 8$; PBS offspring + LPS, $n = 9$; MIA offspring + vehicle, $n = 13$; MIA offspring + LPS, $n = 11$; from 3 independent experiments. Scale bar, 200 µm. **b**, **c**, Representative images (**b**) and quantification (**c**) of FOS (green) and NeuN (red) colabelled cells within the S1DZ of PBS and MIA offspring. PBS offspring, $n = 4$; MIA offspring, $n = 3$; from 1 independent experiment. Scale bar, 50 µm. **d**, **e**, Representative images (**d**) and quantification (**e**) of FOS expression in a series of cortical regions and after injection of vehicle or LPS in PBS or MIA offspring. Sections are stained for FOS (green) and DAPI (blue). Scale bars, 200 µm. For S1BF, M2, M1 and AuD: PBS offspring + vehicle, $n = 8$; PBS offspring + LPS, $n = 9$; MIA offspring + vehicle, $n = 13$; MIA offspring + LPS, $n = 11$. For mPFC: PBS offspring + vehicle, $n = 8$; PBS offspring + LPS, $n = 9$; MIA offspring + vehicle, $n = 12$; MIA offspring + LPS, $n = 11$. For V1: PBS offspring + vehicle, $n = 7$; PBS offspring + LPS, $n = 8$; MIA offspring + vehicle, $n = 12$; MIA offspring + LPS, $n = 9$; from 4 independent experiments. All n values refer to the number of mice used. Statistics calculated by unpaired two-tailed t-test (**c**) or two-way ANOVA with Tukey's post-hoc test (**e**). Graphs are mean ± s.e.m.

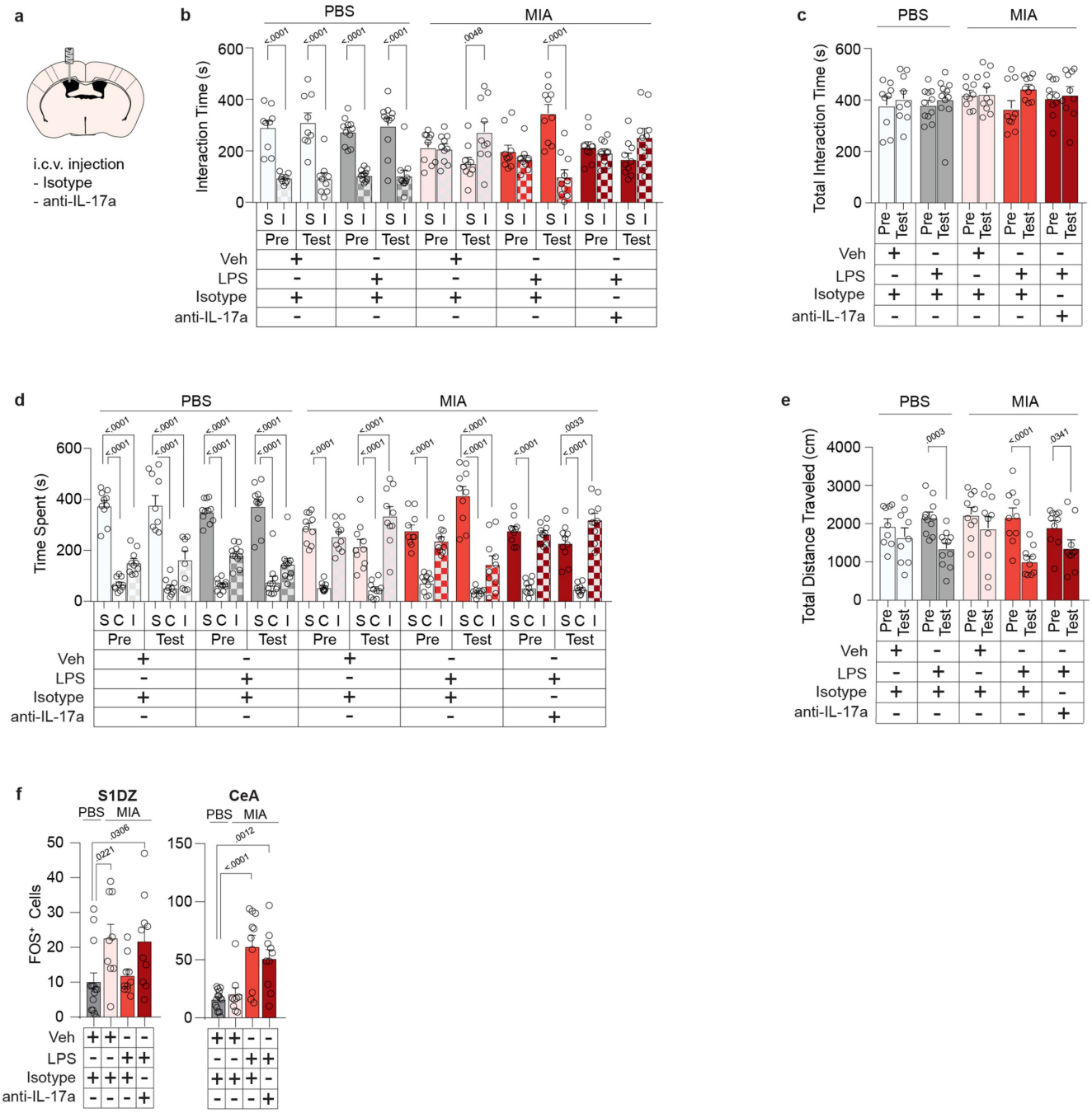


Extended Data Fig. 7 | Further behavioural analyses of S1DZ optical-inhibition-mediated rescue of sociability in monogenic mutant mice. **a**, Quantification of FOS-expressing cells in the S1DZ of monogenic mutant mice. Wild type, $n = 6$; *Cntnap2* mutant, $n = 21$; *Fmr1* mutant, $n = 17$; *Shank3* mutant, $n = 15$; from 5 independent experiments. **b**, Correlation of FOS expression in the S1DZ with severity of sociability deficits across monogenic mutant mice. *Cntnap* mutant, $n = 21$; *Fmr1* mutant, $n = 17$; *Shank3* mutant, $n = 15$; from 4 independent experiments. Black solid lines represent regression line; grey lines indicate 90% confidence intervals. **c**, Individual data for experiments in Fig. 2f. **d**–**f**, Data are expressed as time spent investigating social (S) versus inanimate (I) objects (**d**), total interaction time (**e**), and distance travelled (**f**) during the three-chambered sociability experiments described in Fig. 2f. Wild type + EYFP, $n = 7$; wild type + NpHR, $n = 8$; *Cntnap2* mutant, + EYFP, $n = 11$; *Cntnap2* mutant + NpHR, $n = 9$; *Fmr1* mutant + EYFP, $n = 8$; *Fmr1* mutant + NpHR, $n = 12$; *Shank3* mutant + EYFP, $n = 8$; *Shank3* mutant + NpHR, $n = 10$; from 6 independent experiments. All n values refer to the number of mice used. Statistics calculated by one-way ANOVA with Dunnett's post-hoc test (**a**), linear regression (**b**), one-way repeated-measures ANOVA with Dunnet's post-hoc test (**c**), two-way ANOVA with Sidak's post-hoc test (**d**) or two-way repeated-measures ANOVA with Sidak's post-hoc test (**e**, **f**). Graphs are mean ± s.e.m.

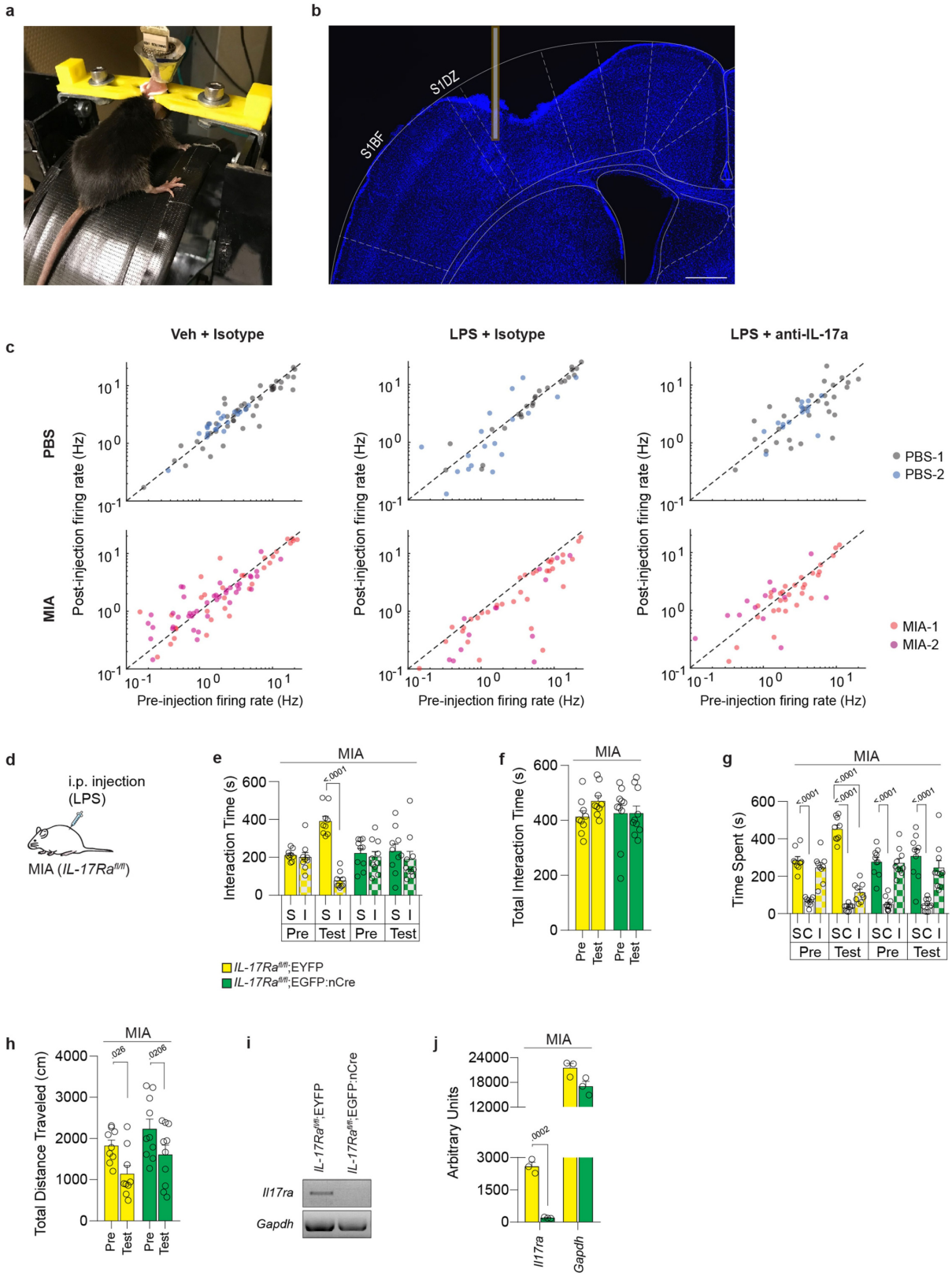


Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | *Il17ra* expression in the S1DZ of PBS and MIA offspring and further behavioural analyses of S1DZ IL-17a rescue of sociability in MIA offspring and monogenic mutant mice. a, Representative images of *Il17ra* expression in the S1DZ of PBS and MIA offspring. Scale bar, 1 mm. **b**, Quantification of *Il17ra* expression within the S1DZ of MIA offspring according to cortical layer. $n = 6$, from 2 independent experiments. **c**, Quantification of overall *Il17ra* expression in the S1DZ of PBS and MIA offspring. PBS offspring, $n = 8$; MIA offspring, $n = 6$; from 2 independent experiments. **d**–**h**, Further behavioural analyses of experiments described in Fig. 3f. Time spent investigating social (S) versus inanimate (I) objects (**e**), total interaction time (**f**), time spent in social (S), centre (C) or inanimate (I) chamber (**g**), and distance travelled (**h**). PBS offspring + vehicle, $n = 11$; PBS offspring + IL-17a, $n = 12$; MIA offspring + vehicle, $n = 14$; MIA offspring + IL-17a, $n = 10$; wild type + vehicle, $n = 11$; wild type + IL-17a, $n = 11$; *Cntnap2* mutant + vehicle, $n = 8$; *Cntnap2* mutant + IL-17a, $n = 10$; *Fmr1* mutant + vehicle, $n = 9$; *Fmr1* mutant + IL-17a, $n = 11$; from 6 independent experiments. All n values refer to the number of mice used. Statistics calculated by unpaired two-tailed t-test (**c**), two-way ANOVA with Sidak's (**e**) or Dunnett's (**g**) post-hoc tests, or two-way repeated-measures ANOVA with Sidak's post-hoc test (**f**, **h**). Graphs are mean ± s.e.m.



Extended Data Fig. 9 | IL-17a is necessary for LPS-induced behavioural rescue and reduction of FOS expression in MIA offspring. a–**e**, Further behavioural analyses of experiments described in Fig. 4a. Time spent investigating social (S) versus inanimate (I) objects (**b**), total interaction time (**c**), time spent in social (S), centre (C) or inanimate (I) chamber (**d**), and distance travelled (**e**). PBS offspring + vehicle + isotype, $n = 9$; PBS offspring + LPS + isotype, $n = 11$; MIA offspring + vehicle + isotype, $n = 10$; MIA offspring + LPS + isotype, $n = 10$; MIA offspring + LPS + anti-IL-17a, $n = 10$; from 7 independent experiments. **f**, Quantification of FOS-expressing cells in the S1DZ and CeA after injection of vehicle or LPS in MIA offspring pre-treated intracerebroventricularly with isotype-control antibody or blocking antibody against IL-17a (anti-IL-17a). PBS offspring + vehicle + isotype, $n = 14$; MIA offspring + vehicle + isotype, $n = 9$; MIA offspring + LPS + isotype, $n = 10$; MIA offspring + LPS + anti-IL-17a, $n = 10$; from 4 independent experiments. All n values refer to the number of mice used. Statistics calculated by two-way ANOVA with Sidak's (**b**) or Dunnett's (**d**) post-hoc tests, two-way repeated-measures ANOVA with Sidak's post-hoc tests (**c**, **e**) or one-way ANOVA with Dunnett's post-hoc test (**f**). Graphs are mean ± s.e.m.



Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | Further analyses of the necessity of IL-17a for the LPS-induced reduction of firing rate in the S1DZ, and the necessity of S1DZ IL-17Ra expression for the LPS-induced rescue of sociability deficits in MIA offspring. a–**c**, Further analyses for experiments described in Fig. 4b–d. **a**, Example of a head-fixed mouse on the running wheel used during single-unit recording. **b**, Representative image of a tetrode placement in the S1DZ. Scale bar, 500 µm. **c**, Firing rate for individual cells before and 4 h after injection of vehicle or LPS in PBS and MIA offspring pre-treated with isotype-control antibody or blocking antibody against IL-17a (anti-IL-17a). **d**–**h**, Further analyses for experiments described in Fig. 4e, f. Time spent investigating social (S) versus inanimate (I) objects (**e**), total interaction time (**f**), time spent in social (S), centre (C) or inanimate (I) chambers (**g**), and distance travelled (**h**). *Il-17ra*$^{fl/fl}$*;EYFP*, $n = 9$; *Il-17ra*$^{fl/fl}$*;EGFP:nCre*, $n = 10$; from 5 independent experiments. **i**, **j**, Representative images (**i**) and corresponding quantification (**j**) of *Il17ra* and *Gapdh* amplicon following PCR using cDNA derived from cells isolated from the cortical region centred on S1DZ of *Il-17ra*$^{fl/fl}$;EYFP and *Il-17ra*$^{fl/fl}$*;EGFP:nCre* mice. $n = 3$ for both groups; from 1 experiment. All n values refer to the number of mice used. Statistics calculated by two-way ANOVA with Sidak's post-hoc test (**e**, **g**), two-way repeated-measures ANOVA with Sidak's post-hoc tests, (**f**, **h**) or unpaired two-tailed t-test (**j**). Graphs are mean ± s.e.m.

nature research

Corresponding author(s): Gloria B. Choi

Last updated by author(s): 10/17/2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Behavioral experiments were recorded using Ethovision XT. IL-17R knockout and c-Fos+ cell quantification were performed using FIJI (ImageJ 1.49m). qPCR data for SFB quantification were collected using Lightcylcler R 96 SW1.1. In situ and immunohistochemistry automated quantification was performed using QuPath. Single unit recording data were acquired using a Neuralynx multiplexing digital recording system. Spike sorting for single-unit recording data was performed using MountainSort and the MClust toolbox.
Data analysis	Data were analyzed using Graphpad Prism v8.0.1.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Provide your data availability statement here.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Sample size was chosen based on similar previous studies (8,15)
Data exclusions	Behavioral results from mice with inaccurate targeting of viral infection or cannula implantations were excluded.
Replication	Key experiments were reiterated with representatives from each group with similar observations across iterations.
Randomization	Within each iteration of an experiment, animals were randomly assigned to groups with approximately balanced sample size.
Blinding	Experimenters were blind to subject treatment during data collection and analysis.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	chicken anti-GFP (1:1000, Ab5450, Abcam), rabbit anti-cFos (1:500, ABE457, Millipore), rabbit anti-DsRed (1:1000, 632496, Clontech), mouse anti-NeuN (1:1000, MAB377, Millipore).
Validation	These are all commonly used antibodies. Statements regarding validation can be found at the manufactures website.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	WT mice used for experiments were male C57BL/6 mice. Male IL17Ra KO, IL17Ra(fl/fl), Cntnap2, Fmr1, Shank3, and Vgat-Cre mice in the C57BL/6 background were also used for experiments. Sources for all mice used are reported in the methods.
Wild animals	Study did not involve wild animals.
Field-collected samples	Study did not involve field-collected samples
Ethics oversight	Committee on Animal Care at Massachusetts Institute of Technology provided approval for all experiments.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Somatic inflammatory gene mutations in human ulcerative colitis epithelium

https://doi.org/10.1038/s41586-019-1844-5

Received: 27 February 2019

Accepted: 31 October 2019

Published online: 18 December 2019

Kosaku Nanki[1,2,7], Masayuki Fujii[1,3,7], Mariko Shimokawa[1], Mami Matano[1], Shingo Nishikori[1,4], Shoichi Date[1,4], Ai Takano[1], Kohta Toshimitsu[1,2], Yuki Ohta[1], Sirirat Takahashi[1], Shinya Sugimoto[1,2], Kazuhiro Ishimaru[1,3], Kenta Kawasaki[1,2], Yoko Nagai[4], Ryota Ishii[5], Kosuke Yoshida[1,2], Nobuo Sasaki[1,2], Toshifumi Hibi[2,6], Soichiro Ishihara[3], Takanori Kanai[2] & Toshiro Sato[1,2]*

With ageing, normal human tissues experience an expansion of somatic clones that carry cancer mutations[1–7]. However, whether such clonal expansion exists in the non-neoplastic intestine remains unknown. Here, using whole-exome sequencing data from 76 clonal human colon organoids, we identify a unique pattern of somatic mutagenesis in the inflamed epithelium of patients with ulcerative colitis. The affected epithelium accumulates somatic mutations in multiple genes that are related to IL-17 signalling—including *NFKBIZ*, *ZC3H12A* and *PIGR*, which are genes that are rarely affected in colon cancer. Targeted sequencing validates the pervasive spread of mutations that are related to IL-17 signalling. Unbiased CRISPR-based knockout screening in colon organoids reveals that the mutations confer resistance to the pro-apoptotic response that is induced by IL-17A. Some of these genetic mutations are known to exacerbate experimental colitis in mice[8–11], and somatic mutagenesis in human colon epithelium may be causally linked to the inflammatory process. Our findings highlight a genetic landscape that adapts to a hostile microenvironment, and demonstrate its potential contribution to the pathogenesis of ulcerative colitis.

Most tissues in our body constantly self-renew to maintain homeostasis. During repetitive cycles of self-renewal, somatic cells inevitably acquire genetic mutations[6]. Sequencing efforts have identified such somatic mutations in normal tissues at a clonal or near-clonal level[1–7]. Notably, these mutated genes often overlap with the driver genes of tumours that originate from the same tissues[6]. Genetic perturbation studies also support the concept that the acquisition of tissue-specific somatic mutations, or 'somatic evolution', confers cell-autonomous self-renewal capability and contributes to tumorigenesis. Given the importance of adaptation that is driven by the environment in Darwinian evolution, somatic evolution may also occur in a non-cell-autonomous manner as a consequence of continuous pressure from the hostile tissue environment. However, the biological intractability of mutant clones in normal tissues renders this question challenging to address. Recurrent mutations have only been captured as sequencing snapshots, and deriving living mutant clones has remained difficult owing to their rarity.

Ulcerative colitis (UC) is an idiopathic chronic inflammatory bowel disease that is characterized by continuous inflammation starting from the rectum. Long-standing inflammation predisposes patients with UC to colorectal cancer, which suggests that somatic evolution occurs in UC. Sporadic emergence of *TP53*-mutant clones in the non-neoplastic UC epithelium supports this hypothesis[12]. However, whether the UC epithelium reorganizes its genetic landscape in parallel with chronic inflammatory exposure has not yet been investigated. To resolve this issue, we here obtain a panoramic view of the genetic mutations in the UC epithelium using an organoid-based analysis of clonal somatic mutations.

Somatic mutations in the UC epithelium

To determine the mutagenic effect of chronic inflammation on normal colorectal epithelium, we established non-neoplastic colorectal organoids from patients with UC, including patients with colitis-associated neoplasia (CAN), and from healthy control individuals. We prioritized organoids that were derived from patients with left-sided UC (that is, UC that involves the colonic segment distal to the splenic flexure) to obtain uninflamed (UC^{uninf}) and inflamed (UC^{inf}) tissue-derived organoids from the same donors. Overall, 170 normal organoid lines representing 55 patients with UC and 16 healthy control individuals (Fig. 1, Extended Data Fig. 1a, b, Extended Data Table 1) were generated. Out of 55 patients with UC, 26 had concomitant CAN. We also established 18 CAN organoid lines. To capture inflammation-related mutations at the clonal level, we subsequently expanded clonal organoids from single cells from patients with UC and healthy control individuals, and subjected the organoids to whole-exome sequencing. The number of single nucleotide variants (SNVs) in the epithelium from healthy control individuals gradually increased with age (0.019 per Mb per year; 95% confidence interval 0.004–0.033; $P = 0.017$), in agreement with a previous report[7] (Fig. 1b). Because the mutational burden in tissues affected by UC was expected to reflect cumulative inflammation rather than disease activity

[1]Department of Organoid Medicine, Keio University School of Medicine, Tokyo, Japan. [2]Department of Gastroenterology, Keio University School of Medicine, Tokyo, Japan. [3]Department of Surgical Oncology, The University of Tokyo, Tokyo, Japan. [4]Fujii Memorial Research Institute, Otsuka Pharmaceutical Company, Shiga, Japan. [5]Biostatistics Unit, Clinical and Translational Research Center, Keio University Hospital, Tokyo, Japan. [6]Center for Advanced IBD Research and Treatment, Kitasato Institute Hospital, Tokyo, Japan. [7]These authors contributed equally: Kosaku Nanki, Masayuki Fujii. *e-mail: t.sato@a7.keio.jp

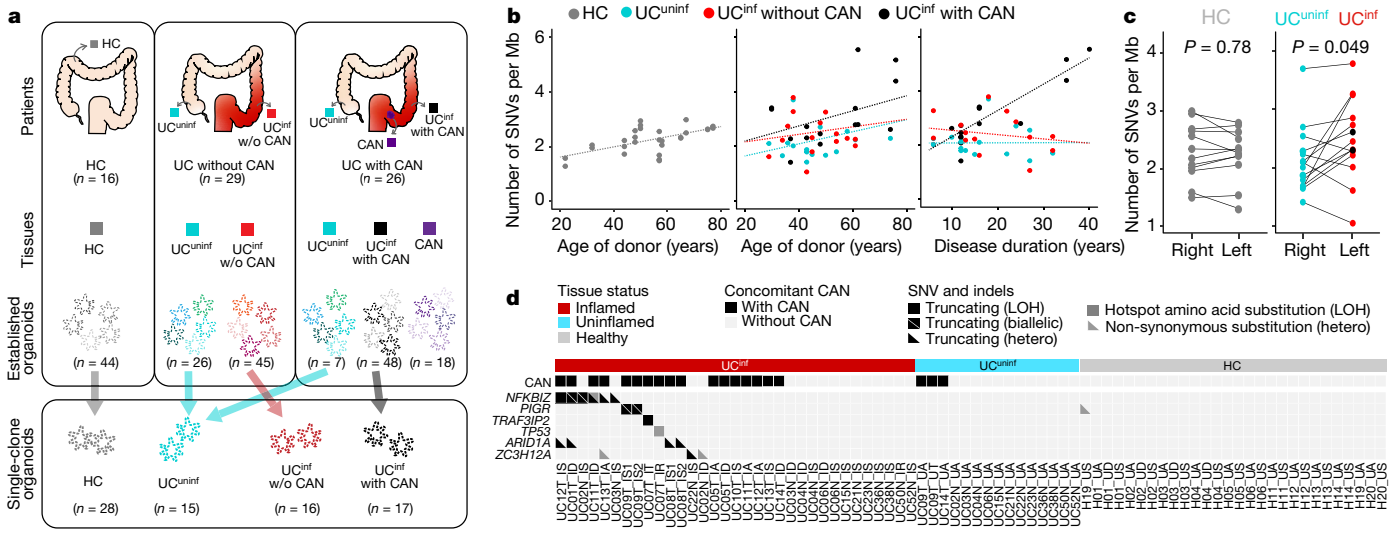


Fig. 1 | Somatic mutations in UC epithelium. a, Strategy for clonal detection of somatic mutations in UC epithelium. Specimens were obtained from healthy control individuals (HC), patients with UC without CAN and patients with UC with CAN. The *n* numbers in the top row top refer to the number of patients; those in the bottom two rows refer to the number of organoid lines. **b**, The relationship between the number of SNVs and age (left, middle) or duration of disease (right) in healthy control ($n = 28$), UC^{uninf} ($n = 15$), UC^{inf} without CAN ($n = 15$) and UC^{inf} with CAN ($n = 15$) samples. Dotted lines show linear regression as determined by the linear mixed model. **c**, Intra-patient comparison of numbers of SNVs between the left and right colons in healthy control individuals ($n = 12$) and patients with left-sided UC ($n = 14$). *P* values determined by two-sided Wilcoxon signed-rank test. **d**, Genetic mutations identified in organoid clones. UC organoids that were derived from patients with CAN are labelled in black. Hetero, heterozygous.

at the point of sampling, we divided UC^{inf} organoids on the basis of CAN concomitance to designate organoids with a history of long-term active inflammation. The rate of mutation did not significantly vary among the groups in a linear mixed model that incorporated age and disease status as fixed effects (Fig. 1b). The number of SNVs was also not correlated with endoscopic severity at the time of sampling. We next assessed the effect of disease duration, as this factor varies among patients with UC. The SNV count in UC^{inf} organoids that were derived from patients with CAN was positively correlated with disease duration (0.10 per Mb per year; 95% confidence interval 0.06–0.14; $P = 0.0017$) (Fig. 1b). This mutation rate was significantly higher than that in UC^{uninf} organoids, even if the interaction between age and disease duration was considered ($P = 0.0032$). These results suggest that chronic inflammation induced genetic mutations independently of ageing, but that other factors not related to inflammation—such as patient medication and the inherent genetic mutability of the host—may have confounded the mutational loads. To exclude such potential bias, we compared SNV counts between paired UC^{inf} and UC^{uninf} organoids that were derived from patients with left-sided UC. This intra-patient comparison demonstrated that the number of SNVs in UC^{inf} organoids was nominally higher than that in UC^{uninf} organoids, whereas SNV counts in organoids from healthy control individuals were comparable between the right and left colons (Fig. 1c).

We next determined whether UC inflammation is related to specific mutational signatures[13]. Regardless of inflammation, the organoids acquired ageing-associated nucleotide transitions (Extended Data Fig. 2a). Despite being derived from a non-neoplastic tissue, UC08 organoids contained an $MLH1^{T117M}$ mutation and showed mutational signatures that are associated with a deficiency in mismatch repair. UC10T_IS organoids also had APOBEC mutational signatures (Extended Data Fig. 2b). Apart from these lines, an association between specific mutational signatures and inflammation was not evident (Extended Data Fig. 2b). Collectively, these results suggest that UC inflammation produces a modest increase in the overall number of mutations, without a bias towards specific mutational signatures.

Genomic landscape of the UC epithelium

To investigate whether functional genetic mutations were enriched in UC^{inf} organoids, we focused on truncating mutations, including nonsense, frame-shift and splice site mutations. In contrast to healthy organoids, UC^{inf} organoids acquired recurrent truncating mutations in two genes that are associated with mucosal immunity—*NFKBIZ* and *PIGR* (Fig. 1d). Five UC^{inf} organoid clones acquired biallelic *NFKBIZ* or *PIGR* truncating mutations, which highlights the inflammation-specific selection of these mutations (Fig. 1d). The gross number of *NFKBIZ* truncating mutations in UC^{inf} organoids was significantly higher compared to other truncating mutations, even in this small sample size (Extended Data Fig. 2c). Furthermore, SNVs in *NFKBIZ* and *PIGR* showed high dN/dS ratios (the ratio of non-synonymous to synonymous substitutions)[14], demonstrating a bias towards non-synonymous, protein-damaging mutations over neutral synonymous mutations (Extended Data Fig. 2d). In two UC^{inf} organoids, truncating mutations in *NFKBIZ* and *TRAF3IP2* were accompanied by loss of heterozygosity (LOH; Fig. 1d). Whole-exome sequencing data also identified LOH in 7 out of 31 UC^{inf} organoids, which predominantly spanned recurrently mutated loci: chromosomes 1q (*PIGR*), 3q (*NFKBIZ*) and 6q (*TRAF3IP2*) (Extended Data Fig. 3a). A DNA copy-number analysis confirmed predicted chromosomal losses in two out of seven UC^{inf} organoids with LOH (Extended Data Fig. 3b). These two clones exhibited aneuploidy, whereas LOH regions in the other five euploid lines showed copy-neutral LOH (CN-LOH) (Extended Data Fig. 3b). One UC^{inf} line (UC12T_IA), which had CN-LOH in the *PIGR* locus without a somatic mutation in the *PIGR* coding region, lacked expression of the PIGR protein—emphasizing the role of LOH in genetic disruption (Extended Data Fig. 3c). We also detected focal deletions at fragile sites exclusively in UC^{inf} organoids (Extended Data Fig. 3d). These results indicate that the UC epithelium is prone to genomic instability, and that this eventually contributes to the formation of a unique genomic landscape.

Recurrent IL-17 pathway mutations in UC

We next investigated biological pathways that are affected by UC-related mutations. *NFKBIZ* encodes IκBζ, which is upregulated by NF-κB signal activation and enhances the activation of NF-κB by regulating positive feedback[15,16]. PIGR is a cell-surface molecule that regulates transcytosis and the secretion of dimeric immunoglobulin A (IgA)[17]. Previous studies have revealed that PIGR is upregulated by IL-17-related

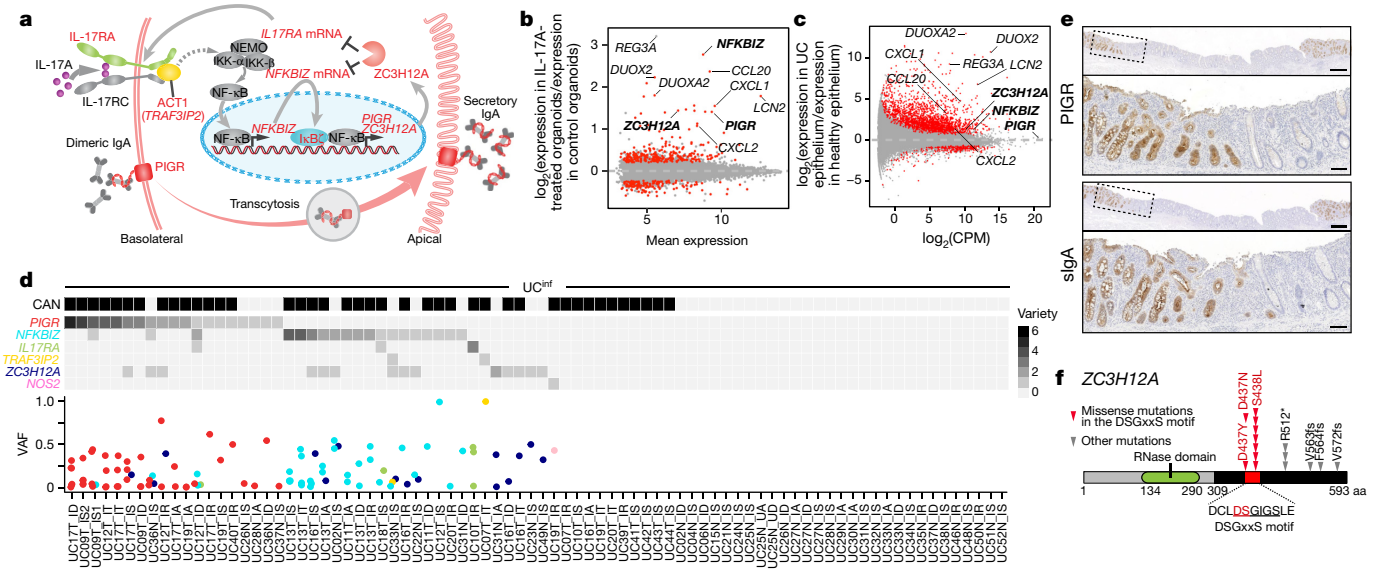


Fig. 2 | Targeted sequencing of bulk UC organoids. a, The IL-17 signalling pathway. Genes that are mutated in UC are in red. **b**, IL-17A-responsive genes in normal colon organoids ($n = 3$). Genes with $P < 0.01$ (empirical Bayes method) are shown in red. Bold text indicates genes that are mutated in UC. **c**, Genes that are differentially expressed between purified UC ($n = 11$) and healthy ($n = 11$) epithelium (E-MTAB-5464). Genes with adjusted $P < 0.01$ (exact negative binomial test followed by the Benjamin–Hochberg procedure) are shown in red. Bold text indicates genes that are mutated in UC. CPM, counts per million. **d**, Genetic mutations detected in UC^{inf} organoids by targeted sequencing. Top, for each gene, the number of different variants is shown by density. Bottom, dots represent the VAF of each variant. Genes are shown in different colours as specified on the left. **e**, PIGR (top) and secretory IgA (sIgA) (bottom) immunostaining of a UC^{inf} tissue. A *PIGR*-mutant organoid line (UC12T_IT) was derived from the adjacent region. Magnified areas show boundaries between regions in which PIGR expression is intact and regions in which PIGR expression is lost. Representative result from $n = 8$ patients analysed. Scale bars, 500 µm (low magnification); 100 µm (high magnification). **f**, *ZC3H12A* mutations identified by targeted sequencing. All mutations in UC organoids were located in the last exon (black). fs, frameshift; *, nonsense mutation.

activation of NF-κB[18,19]. TRAF3IP2 (also known as ACT1) serves as an adaptor of the IL-17 receptor and is essential for downstream activation of NF-κB signalling[20]. Recurrently mutated genes in the inflamed UC epithelium therefore appear to converge on the IL-17–NF-κB signalling pathway (Fig. 2a). In support of this, IL-17A stimulation upregulated *NFKBIZ*, *PIGR* and other canonical targets of IL-17A, including *CXCL1* and *LCN2*, in colonic organoids (Fig. 2b). A reanalysis of published datasets revealed upregulation of *NFKBIZ* mRNA and other IL-17A target genes in tissues with UC (Fig. 2c, Extended Data Fig. 4a, b). Together, these results suggest that IL-17 signal activation upregulates the expression of IκBζ and PIGR in the UC epithelium.

To determine whether IκBζ is a critical regulator of IL-17 signal activation in the colonic epithelium, we treated wild-type, *NFKBIZ*-mutant and *TRAF3IP2*-mutant UC^{inf} organoids with IL-17A. PIGR was not upregulated in *NFKBIZ*- and *TRAF3IP2*-mutant organoids that were treated with IL-17A (Extended Data Fig. 4c–h), highlighting the convergence of mutations in the IL-17 and IκBζ signalling pathways during UC-related inflammation. Of note, mutations in *NFKBIZ* and *PIGR* were rare or heterozygous in CANs in our cohort (Extended Data Fig. 4i)—as well as in previous exome-sequencing analyses[21–24]—which suggests that alteration of these genes is unlikely to initiate colitis-associated tumorigenesis.

The mutational spectrum in UC epithelium

To assess the prevalence of UC-associated mutations, we subjected short-term cultures of epithelium-enriched polyclonal UC^{inf}, UC^{uninf} and control organoids to targeted sequencing of 37 genes that comprise components of the IL-17 signalling pathway and sporadic drivers of colon cancer (Fig. 2a, Supplementary Table 2). The analysis confirmed that *NFKBIZ* and *PIGR* mutations were frequent in UC^{inf} organoids (Fig. 2d, Supplementary Table 2). Variant-allele frequencies (VAFs) were similar between the original epithelium tissues and the organoids, which suggests that selection bias was minimal during organoid derivation (Extended Data Fig. 5a). In three UC^{inf} organoids, we identified five *IL17RA* mutations that were not detected by whole-exome sequencing. Multiple disparate truncating mutations often co-existed in a tiny biopsy sample, with *PIGR*, *ARID1A* and *NFKBIZ* mutations being prevalent. The high dN/dS ratios of these mutations indicated subsequent functional effects (Extended Data Fig. 5b). Targeted sequencing also detected UC-associated mutations in UC^{uninf} organoids, but the medical records of the respective patients revealed a past history of inflammation in most of the regions that were sampled—suggesting that previous inflammation has a role in the selection of UC-associated mutations (Extended Data Fig. 5c).

On the basis of the high frequency of mutations in *PIGR*, we next examined the expression of PIGR in the colonic tissues adjacent to the sampled sites. PIGR was ubiquitously expressed in healthy and UC^{uninf} epithelium, whereas UC^{inf} tissues exhibited a focal loss of PIGR that was attributable to *PIGR* mutations (Fig. 2e, Extended Data Fig. 5d). The concomitant depletion of secretory IgA in these regions reinforced the functional loss of PIGR (Fig. 2e, Extended Data Fig. 5d). Notably, several studies from the 1970s and 1980s documented idiopathic loss of epithelial IgA and PIGR in patients with UC[25–27]. The frequent somatic *PIGR* mutations could constitute a previously unknown mechanism for the sporadic disappearance of the IgA system in the UC epithelium.

Target sequencing also identified a *ZC3H12A*^{S438L} mutation in seven UC^{inf} organoids. ZC3H12A, which is also known as Regnase-1, is an IL-17-responsive RNA-binding protein that dismantles the mRNA of inflammation-related genes, including *NFKBIZ* and *IL17RA*[28]. Phosphorylation of serine residues in its DSGXXS phospho-degron motif, in which S438 is situated (Fig. 2f), recruits the ubiquitin machinery to degrade the ZC3H12A protein. As the S438L mutation blocks this phosphorylation process[29], this acts as a gain-of-function mutation in *ZC3H12A* and results in the stabilization of the protein. One UC^{inf} organoid line carried a *ZC3H12A*^{D437Y} mutation that affects the DSGXXS motif (Fig. 2f), and this organoid line exhibited a lower response to IL-17A than

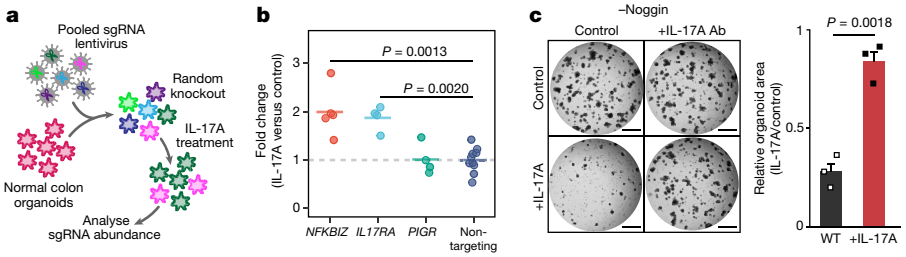


Fig. 3 | Phenotypic characterization of UC-related mutations. a, Strategy for screening for genetic mutations that confer IL-17A resistance. **b**, Enrichment of sgRNA-targeted organoids after four weeks of treatment with IL-17A. Fold-change values relative to untreated organoids are shown. Each dot corresponds to one sgRNA sequence. Bars represent mean fold-change values. Data are representative of $n = 2$ biologically independent experiments with similar results. *P* values determined by two-sided Mann–Whitney *U* test. **c**, Sensitization of organoids to IL-17A by withdrawal of Noggin, and its reversal by pre-treating organoids with a neutralizing anti-IL-17A antibody (Ab). Right, dots show the relative area of organoids in each well. WT, wild type. Data are representative of $n = 3$ biologically independent experiments with similar results, and are shown as mean + s.e.m. *P* value determined by two-sided Welch's *t*-test. Scale bars, 1 mm.

did patient-matched wild-type *ZC3H12A* organoids (Extended Data Fig. 6a, b). We also generated a *ZC3H12A*^{D437Y} knock-in (KI) line of organoids (*ZC3H12A*$^{D437Y\text{-}KI}$) and found that this showed a similar response to IL-17A (Extended Data Fig. 6c–e), verifying the gain of function that is conferred by the D437Y mutation. These results suggest that gain-of-function hotspot mutations, in analogy to oncogenic mutations in cancers, can occur in non-dysplastic human epithelium.

To exclude the potential overrepresentation of specific mutations owing to our use of multiple clones derived from the same patients, we reanalysed the targeted sequencing data by selecting the most-distal sample for each patient. This per-patient analysis indicated that around 65% of the patients with UC and concomitant CAN had at least one of the IL-17-signalling-related mutations in an area of epithelium of less than 5 mm^2 (Extended Data Fig. 6f).

Given that *NFKBIZ*, *ZC3H12A* and *PIGR* could be upregulated by cytokines other than IL-17A, we comprehensively examined mutations in cytokine and pathogen-associated molecular pattern (PAMP) receptor genes by targeted sequencing. Although none of the 76 selected cytokine or PAMP receptor genes was mutated in more than two samples, several truncating mutations were identified (Extended Data Fig. 7a). Of the genes with truncating mutations, *OSMR*, *IL15RA* and *TLR3* were found to be expressed at substantial levels in the colonic epithelium[30]. Nonetheless, genetic analysis on a larger scale will be required to confirm their enrichment during inflammation.

Targeted sequencing also detected sporadic colon cancer-related mutations in both UC and control epithelium, with higher incidence in the UC epithelium (Extended Data Fig. 7b, c). Cancer-associated mutations that were identified in UC^{inf} organoids were also present in CAC organoids, which suggests that these mutations have a selection advantage under conditions of chronic inflammation (Extended Data Fig. 7d).

IL-17A cytotoxicity in colonic epithelium

The frequent mutations that occur in the IL-17 signalling pathway suggested the presence of a mutation-driven selection bias during a prolonged period of inflammation. To determine whether IL-17 signalling elicits cytotoxic effects in the colonic epithelium, we treated colonic organoids with IL-17A. Contrary to our expectation, a week-long treatment with IL-17A marginally affected organoid growth, which hampered the functional assessment of UC-associated mutations. We thus turned to an agnostic CRISPR-based knockout screening (Fig. 3a). This approach sensitively identified genes the loss of which confers a survival advantage during IL-17A selection, as demonstrated by an enrichment of *IL17RA* and *NFKBIZ* single-guide (sg)RNAs. (Fig. 3b, Extended Data Fig. 8a). Of note, *PIGR* sgRNAs were not enriched, which suggests that growth selection by *PIGR* mutations was not mediated directly by IL-17A cytotoxicity but requires an IgA-related immune response.

We next asked whether specific medium conditions evoke IL-17A-mediated organoid death. Screening of growth factor conditions demonstrated that Noggin deficiency sensitizes organoids to treatment with IL-17A (Extended Data Fig. 8b). An anti-IL-17A neutralizing antibody or genetic disruption of *IL17RA* counteracted this cytotoxic response (Fig. 3c, Extended Data Fig. 8c, d). Both *NFKBIZ* knockout (KO) and patient-derived *NFKBIZ*-mutant organoid lines exhibited resistance to treatment with IL-17A (Extended Data Fig. 8c, e, f). Furthermore, *TRAF3IP2*^{KO} and *ZC3H12A*^{D437Y} organoids showed a similar response (Extended Data Fig. 8c, d, g). *TP53*^{KO} organoids also tolerated IL-17A stimulation (Extended Data Fig. 8h), which may account for the emergence of *TP53*-mutant clones in non-neoplastic UC epithelium and nascent CANs. These results together suggest that IL-17A is detrimental for the epithelial lining, and that clones with IL-17A pathway mutations are able to avert IL-17A-mediated cytotoxicity and thereby selectively expand in tissue that is affected by UC.

Mechanism of IL-17A-induced cytotoxicity

An increase in cells that express cleaved caspase-3 in wild-type, but not *NFKBIZ*^{KO}, organoids as a result of treatment with IL-17A indicated an activation of the apoptotic pathway (Fig. 4a, Extended Data Fig. 9a). Transcriptome analysis revealed a biased induction of *NOS2* (iNOS, inducible NO synthase) in a Noggin-deficient condition (Fig. 4c), suggesting that *NOS2* has a role in IL-17A-induced apoptosis. Treatment with iNOS inhibitors suppressed IL-17A-induced growth arrest (Fig. 4d). IL-17A stimulation failed to induce *NOS2* expression in *NFKBIZ*^{KO}, *TRAF3IP2*^{KO} and *IL17RA*^{KO} organoids (Extended Data Fig. 9b), consistent with their resistance to IL-17A. Of note, targeted sequencing detected one *NOS2* truncating mutation in UC^{inf} organoids (Fig. 2d). In agreement with previous studies[31–33], iNOS was highly expressed at the mucosal surface of UC^{inf} tissue (where BMP signalling is active), but its expression was scarce in healthy tissue (Fig. 4f). Together, these data suggest that in tissue with UC, IL-17A transmits iNOS- and NO-dependent pro-apoptotic signals to the superficial layer of the epithelium and subsequently drives the selective expansion of IL-17A-resistant mutant clones (Fig. 4g).

Discussion

In contrast to the pervasive spread of cancer mutations in the haematopoietic system[1,2], skin[5] and the epithelium of the oesophagus[3,4], the UC epithelium accrues tumour-irrelevant mutations that are presumed to facilitate selective expansion of these cells under inflammatory conditions. Using organoid culture, we pinpointed IL-17 signalling as one of the key pathways that drives this somatic evolution. IL-17A elicited a pro-apoptotic response in normal colonic organoids, and loss-of-function mutations in positive regulators of IL-17 signalling—namely

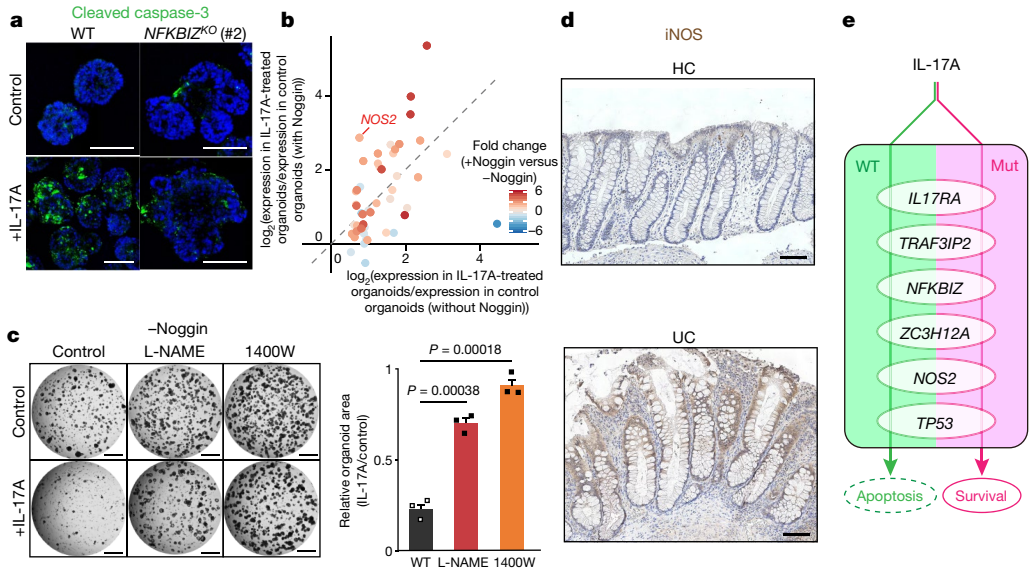


Fig. 4 | iNOS mediates IL-17A-induced apoptosis. a, Expression of cleaved caspase-3 in IL-17A-treated wild-type and *NFKBIZ*^{KO} organoids. Images are representative of $n = 2$ technical replicates with similar results. Scale bars, 100 µm. **b**, Expression of IL-17A target genes after IL-17A stimulation in the absence or presence of Noggin. The colour scale shows the fold change in gene expression that is induced by Noggin withdrawal in the presence of IL-17A. **c**, The iNOS inhibitors L-NAME and 1400W attenuate IL-17A-induced cytotoxicity. Right, dots show the relative area of organoids in each well. Data are shown as mean + s.e.m. *P* value determined by two-sided Welch's *t*-test. Scale bars, 1 mm. **d**, iNOS immunostaining in healthy control and UC^{inf} tissues. Representative images from $n = 3$ (HC) and $n = 6$ (UC) samples with similar staining patterns. Scale bars, 100 µm. **e**, Proposed role of UC-associated somatic mutations in resistance to IL-17A-induced apoptosis.

NFKBIZ, *IL17RA* and *TRAF3IP2*—were enriched in the UC epithelium. Unlike these mutations, alterations in *PIGR* did not translate into protection from IL-17A-mediated cytotoxicity. Although IL-17A induced the expression of PIGR in organoids, high levels of PIGR were present in the colonic epithelium independently of IL-17A stimulation, suggesting that there is room for debate on the convergence of IL-17 signalling mutations. The system that we used for organoid culture lacks IgA and related components of the immune system, which makes it challenging to investigate the effect of *PIGR* mutations during the selection process.

By contrast with the reduction of the IL-17A-induced pro-apoptotic response that was observed here in human organoids with UC-related mutations, loss-of-function mutations of *Il17ra*[8], *Pigr*[9] or *Traf3ip2*[10] (and gain-of-function of *Zc3h12a*[11]) exacerbate experimental colitis in genetically engineered mice. IL-17 signalling was associated with mucosal integrity[10]. Disruption of the IL-17–PIGR pathway provoked dysbiosis and intensified the host immune response[8,9]. Furthermore, a previous clinical trial that aimed to test the efficacy of an anti-IL-17A antibody for the treatment of inflammatory bowel disease was prematurely discontinued as a result of the aggravation of colitis in trial participants[34]. It is thus tempting to speculate that sporadic UC-associated mutations confer a local survival advantage under chronic inflammation, whereas extensive expansion of mutant clones may paradoxically worsen the activity of UC. Notably, some of the somatically mutated genes in the UC epithelium are also associated with susceptibility to UC—including *NFKBIZ* (rs616597), *TRAF3IP2* (rs3851228), *PIGR* (rs3024505) and *NOS2* (rs2945412)[35]—which suggests that altered function of these genes may also contribute to the onset of UC. Further studies are warranted to clarify how these UC-related genes pertain to disease aetiology and pathophysiology on a regional, tissue-wide and systemic level.

In conclusion, our genetic profiling of UC organoids has unveiled a pattern of somatic evolution during chronic tissue inflammation. Our findings provide a blueprint for how clones that are adapted to an inflammatory microenvironment emerge and expand in a normal tissue, and will prompt future research on the role of such clones in the pathology of human diseases.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1844-5.

1. Genovese, G. et al. Clonal hematopoiesis and blood-cancer risk inferred from blood DNA sequence. *N. Engl. J. Med.* **371**, 2477–2487 (2014).
2. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
3. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
4. Martincorena, I. et al. Somatic mutant clones colonize the human esophagus with age. *Science* **362**, 911–917 (2018).
5. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
6. Martincorena, I. & Campbell, P. J. Somatic mutation in cancer and normal cells. *Science* **349**, 1483–1489 (2015).
7. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).
8. Maxwell, J. R. et al. Differential roles for interleukin-23 and interleukin-17 in intestinal immunoregulation. *Immunity* **43**, 739–750 (2015).
9. Murthy, A. K., Dubose, C. N., Banas, J. A., Coalson, J. J. & Arulanandam, B. P. Contribution of polymeric immunoglobulin receptor to regulation of intestinal inflammation in dextran sulfate sodium-induced colitis. *J. Gastroenterol. Hepatol.* **21**, 1372–1380 (2006).
10. Lee, J. S. et al. Interleukin-23-independent IL-17 production regulates intestinal epithelial permeability. *Immunity* **43**, 727–738 (2015).
11. Nagahama, Y. et al. Regnase-1 controls colon epithelial regeneration via regulation of mTOR and purine metabolism. *Proc. Natl Acad. Sci. USA* **115**, 11036–11041 (2018).
12. Brentnall, T. A. et al. Mutations in the *p53* gene: an early marker of neoplastic progression in ulcerative colitis. *Gastroenterology* **107**, 369–378 (1994).
13. Alexandrov, L. B. et al. Signatures of mutational processes in human cancer. *Nature* **500**, 415–421 (2013).
14. Martincorena, I. et al. Universal patterns of selection in cancer and somatic tissues. *Cell* **171**, 1029–1041 (2017).
15. Yamamoto, M. et al. Regulation of Toll/IL-1-receptor-mediated gene expression by the inducible nuclear protein IκBζ. *Nature* **430**, 218–222 (2004).
16. Karlsen, J. R., Borregaard, N. & Cowland, J. B. Induction of neutrophil gelatinase-associated lipocalin expression by co-stimulation with interleukin-17 and tumor necrosis factor-alpha is controlled by IκB-ζ but neither by C/EBP-β nor C/EBP-δ. *J. Biol. Chem.* **285**, 14088–14100 (2010).

17. Kaetzel, C. S., Robinson, J. K., Chintalacharuvu, K. R., Vaerman, J. P. & Lamm, M. E. The polymeric immunoglobulin receptor (secretory component) mediates transport of immune complexes across epithelial cells: a local defense function for IgA. *Proc. Natl Acad. Sci. USA* **88**, 8796–8800 (1991).
18. Kumar, P. et al. Intestinal interleukin-17 receptor signaling mediates reciprocal control of the gut microbiota and autoimmune inflammation. *Immunity* **44**, 659–671 (2016).
19. Cao, A. T., Yao, S., Gong, B., Elson, C. O. & Cong, Y. Th17 cells upregulate polymeric Ig receptor and intestinal IgA and contribute to intestinal homeostasis. *J. Immunol.* **189**, 4666–4673 (2012).
20. Chang, S. H., Park, H. & Dong, C. Act1 adaptor protein is an immediate and essential signaling component of interleukin-17 receptor. *J. Biol. Chem.* **281**, 35603–35607 (2006).
21. Robles, A. I. et al. Whole-exome sequencing analyses of inflammatory bowel disease-associated colorectal cancers. *Gastroenterology* **150**, 931–943 (2016).
22. Fujita, M. et al. Genomic landscape of colitis-associated cancer indicates the impact of chronic inflammation and its stratification by mutations in the Wnt signaling. *Oncotarget* **9**, 969–981 (2018).
23. Baker, A. M. et al. Evolutionary history of human colitis-associated colorectal cancer. *Gut* **68**, 985–995 (2019).
24. Din, S. et al. Mutational analysis identifies therapeutic biomarkers in inflammatory bowel disease-associated colorectal cancers. *Clin. Cancer Res.* **24**, 5133–5142 (2018).
25. Brandtzaeg, P., Baklien, K., Fausa, O. & Hoel, P. S. Immunohistochemical characterization of local immunoglobulin formation in ulcerative colitis. *Gastroenterology* **66**, 1123–1136 (1974).
26. Das, K. M., Erber, W. F. & Rubinstein, A. Immunohistochemical changes in morphologically involved and uninvolved colonic mucosa of patients with idiopathic proctitis. *J. Clin. Invest.* **59**, 379–385 (1977).
27. Uchima, H., Eishi, Y., Takemura, T. & Hirokawa, K. Immunohistochemical studies of ulcerative colitis. With special reference to localization of immunoglobulins, secretory component, and lysozyme in view of suffering periods. *Acta Pathol. Jpn.* **33**, 1183–1196 (1983).
28. Garg, A. V. et al. MCPIP1 endoribonuclease activity negatively regulates interleukin-17-mediated signaling and inflammation. *Immunity* **43**, 475–487 (2015).
29. Iwasaki, H. et al. The IκB kinase complex regulates the stability of cytokine-encoding mRNA induced by TLR-IL-1R by controlling degradation of regnase-1. *Nat. Immunol.* **12**, 1167–1175 (2011).
30. Howell, K. J. et al. DNA methylation and transcription patterns in intestinal epithelial cells from pediatric patients with inflammatory bowel diseases differentiate disease subtypes and associate with outcome. *Gastroenterology* **154**, 585–598 (2018).
31. Kimura, H. et al. Increased expression of an inducible isoform of nitric oxide synthase and the formation of peroxynitrite in colonic mucosa of patients with active ulcerative colitis. *Gut* **42**, 180–187 (1998).
32. Kolios, G., Rooney, N., Murphy, C. T., Robertson, D. A. & Westwick, J. Expression of inducible nitric oxide synthase activity in human colon epithelial cells: modulation by T lymphocyte derived cytokines. *Gut* **43**, 56–63 (1998).
33. Parikh, K. et al. Colonic epithelial cell diversity in health and inflammatory bowel disease. *Nature* **567**, 49–55 (2019).
34. Hueber, W. et al. Secukinumab, a human anti-IL-17A monoclonal antibody, for moderate to severe Crohn's disease: unexpected results of a randomised, double-blind placebo-controlled trial. *Gut* **61**, 1693–1700 (2012).
35. McGovern, D. P., Kugathasan, S. & Cho, J. H. Genetics of inflammatory bowel diseases. *Gastroenterology* **149**, 1163–1176 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Patient enrolment and tissue collection

Human intestinal samples were obtained from healthy control individuals or patients with UC who underwent colonoscopy or surgery at Keio University Hospital or Tokyo University Hospital, with written informed consent under approval of the ethical committees (approval numbers: 20120057, 20130512, 20170170, 20170171 and G3553-(7)). UC was classified into proctitis, left-sided colitis or extensive colitis according to the standard clinical criteria[36], and disease distribution was determined by macroscopic or endoscopic diagnosis at the time of tissue sampling. Whenever possible, multiple mucosal tissues were sampled from differing anatomical segments of the same patient. CAN specimens were obtained in parallel with normal counterparts from patients who were histologically diagnosed with CAN. Normal tissues of patients with CAN were sampled from macroscopically normal regions with gross margins of at least 3 cm from the tumours. Collected tissues were transferred into ice-cold phosphate-buffered saline (PBS) and placed on ice until use. Reference blood samples were obtained from all patients. Patient demographics and sample information are available in Extended Data Table 1.

Crypt isolation and organoid culture

Crypt isolation and organoid culture from samples of human large intestine were carried out as previously described[37,38] following different procedures depending on sample types. For healthy control samples, the stroma was removed using fine scissors and the remaining epithelium was shredded into 1-mm^3 pieces. Epithelial fragments were washed vigorously at least five times with ice-cold PBS until the supernatant was free of debris and were treated with 2.5 mM EDTA (Thermo Fisher Scientific) for 1 h at 4 °C with gentle rocking to release crypts. Liberated crypts were suspended in Matrigel and dispensed onto 48-well plates as 25-µl droplets. After Matrigel polymerization, crypts were overlaid with the medium as described below. Advanced Dulbecco's modified Eagle's medium/F12 (Thermo Fisher Scientific) was supplemented with penicillin and streptomycin (Thermo Fisher Scientific), 10 mM HEPES (Thermo Fisher Scientific), 2 mM GlutaMAX (Thermo Fisher Scientific), 1 × B-27 supplement (Thermo Fisher Scientific), 10 nM gastrin I (Sigma-Aldrich), 1 mM *N*-acetylcysteine (Sigma-Aldrich), 100 ng ml^{-1} recombinant mouse Noggin (PeproTech), 50 ng ml^{-1} recombinant mouse EGF (Thermo Fisher Scientific), 100 ng ml^{-1} recombinant human IGF-1 (BioLegend), 50 ng ml^{-1} recombinant human FGF-basic (FGF-2) (Peprotech), 1 µg ml^{-1} recombinant human R-spondin-1 (R&D), 500 nM A83-01 (Tocris) and 20% Afamin/WNT3A serum-free conditioned medium[39]. For UC^{uninf}, UC^{inf} and CAN samples, fragmented mucosal tissues were digested in Liberase TH Research Grade (Roche) for 1 h at 37 °C. The remaining fragments were further dissociated in TrypLE Express (Thermo Fisher Scientific) for 20 min at 37 °C. For some samples, half of the processed tissue was spared and used for targeted sequencing. Digested tissue was embedded in Matrigel and, regardless of the sample type and inflammation status, maintained in the unified culture medium described above. The medium was refreshed every two days and the organoids were passaged once a week with a split ratio of 1:10 using TrypLE Express. The medium was supplemented with 10 µM Y-27632 (FUJIFILM Wako Pure Chemical Corporation) for two days after establishment and passaging, and 5 µM CHIR99021 (Cayman Chemical) for seven days after establishment.

To evaluate the effect of IL-17A treatment on organoid growth in a Noggin-deficient condition, the organoids were single-cell-passaged and allowed to grow in the absence of Noggin with or without 100 ng ml^{-1} IL-17A. The organoids were cultured for 6–8 days. For iNOS blockade, organoids were treated an IL-17A neutralizing antibody (Secukinumab, 100 nM) from passaging, or with L-NAME (1 mM) or 1400W (500 mM) during IL-17A stimulation. The images of each well were captured using a BZ-X800 digital microscope (Keyence). The organoid area was measured using BX-X800 Analyzer software (Keyence).

Exome sequencing

DNA was extracted from organoids or blood with the QIAamp Blood Mini Kit (Qiagen) and treated with RNase Cocktail (Ambion). The quality of DNA samples was confirmed by agarose gel electrophoresis within the expected range of DNA fragmentation. Three micrograms of genomic DNA was fragmented for whole-exome sequencing analyses as previously described[40]. Whole-exome sequencing libraries were generated by the SureSelect Human All Exon V5 or V6 kit (Agilent) according to the manufacturer's instructions and enriched exon fragments were sequenced by Illumina HiSeq2500 (outsourced to Hokkaido System Science) or HiSeq4000 (outsourced to BGI JAPAN K.K.) in 150-bp paired-end mode. Cleaned fastq files were mapped onto the human reference genome version GRCh37 (hg19) using Burrows-Wheeler Aligner (BWA, v.0.7.8). Generated reads were passed to the Genome Analysis Toolkit (GATK, v.1.6.13). SNVs, insertions and deletions were called using the UnifiedGenotyper tool in GATK. Variants were annotated using SnpEff (v.3). For analysis of numbers of mutations, raw variants were filtered by removing SNVs that were detected in the paired blood samples with VAF > 0.03. The filtered variants in autosomal chromosomes were counted. UC samples with a *MLH1* mutation were excluded from the analysis. UC06_IS were also excluded owing to the use of different versions of exon-capture kits between organoids and reference blood samples. A linear regression model was fitted using the lme function in the R package nlme (v.3.1) to estimate the effect of age, disease duration and group on the number of SNVs. For detection of mutations, raw variants were filtered by removing those registered in the SNV databases (dbSNP build 130 or lower), in the 1000 Genome Project with a frequency of >0.05 or in the Japanese SNP database (the Human Genetic Variation Database (HGVD, release v.2.30). Filtered data were used to calculate dN/dS ratios using the R package dndscv (v.0.0.1.0). The truncating mutations in the filtered data were assessed by Fisher's exact test. Somatic mutations identified by whole-exome sequencing and dndscv results are available in Supplementary Table 1.

Mutational signatures

The nucleotide variants that were used for mutation-number analysis were also subjected to an analysis of mutational signatures. In this analysis, we included organoids with *MLH1* mutations (UC08T_IL1, 2). Using the R package MutationalPatterns (v.1.8.0), the occurrence of six nucleotide changes was counted for each group (UC^{inf}, UC^{uninf}, UC^{inf} without CAN, UC^{inf} with CAN), and the cosine similarity[41] to the COSMIC signatures[13] was calculated. Hierarchical clustering was performed using Euclidean distance and the complete linkage method.

Copy-number analysis

For copy-number analysis, 500 ng of genomic DNA was applied to comparative genomic hybridization (CGH)–SNP microarray (CytoScan HD, Affymetrix), according to the manufacturer's instructions. We used 'apt-copynumber-cyto-ssa' in Affymetrix Power Tools (v.1.19.0) with the normal-diploid-normalization option to obtain probe-level copy number and allele difference. All parameters were set according to the recommendation in user manual. Segmentation analysis was performed with the segmentCGH function in the R Bioconductor package rCGH (v.1.8.1) using $\log_2$-transformed ratio values of copy-number probes.

Gene expression microarray analysis and gene set enrichment analysis

Three independent normal colonic organoid lines that were established previously as counterparts of colon cancer organoids[38] were used for

gene expression microarray analysis. The organoids were cultured from single cells for five days and then treated with or without 100 ng ml^{-1} IL-17A for 24 h. Two of these organoid lines were also used to evaluate the effect of Noggin withdrawal on IL-17A response. In this experiment, the organoids were cultured in Noggin-included or Noggin-deficient conditions for five days and IL-17A was added 24 h before RNA collection. RNA was extracted from organoids using the RNeasy Plus Mini Kit (Qiagen). RNA quality was measured using Bioanalyzer (Agilent), and all samples passed the quality check (RNA integrity number > 7.0). Gene expression profiling was performed using the PrimeView Human Gene Expression Array (Affymetrix) according to the manufacturer's instructions. Raw expression data were normalized using the rma function in the affy R Bioconductor package (v.1.56.0). Linear model fitting and the empirical Bayes method implemented in the R/Bioconductor package limma[42] was used to compare IL-17A-treated and -untreated samples in a pairwise fashion. Forty-nine genes with P values <0.05 and fold-change values >1.5 were selected as in-house IL-17A target genes (Supplementary Table 3). Gene set enrichment analysis was performed using this IL-17A target gene set and public datasets with data on gene expression in UC-affected mucosa (GSE38713[43] and GSE59071[44]). Enrichment of the genes in active colitis samples versus healthy control samples was evaluated using the gene set enrichment analysis module (v.19.0.25) in GenePattern (https://cloud.genepattern.org/gp). Transcriptomic analysis of purified intestinal epithelial cells from patients with UC was performed using a public dataset. The read-count matrix was downloaded from ArrayExpress (study ID: E-MTAB-5464). We examined expression data of sigmoid colon epithelium from 11 patients with UC and 11 healthy control individuals. Exact negative binomial test was performed using the R Bioconductor package edgeR (v.3.24.3).

Targeted sequencing

We used short-term-cultured (fewer than four passages) bulk organoids and fresh epithelium tissues for targeted sequencing. DNA was extracted using the QIAamp Blood Mini Kit (Qiagen) and treated with RNase Cocktail (Ambion). The targeted genes consist of genes that were recurrently mutated in the UC organoid whole-exome sequencing data; IL-17 signalling-related genes (based on the literature); cytokine or PAMP receptor genes; UC risk genes identified in previous genome-wide association studies; and colon cancer driver genes, as listed in Supplementary Table 2. Cytokine and PAMP receptors were extracted from data from the FANTOM5 project[45] and genes with counts per million greater than 1 in at least 50% of sorted UC and control epithelium samples[30] were used for sequencing. Molecular barcode-based high-sensitivity variant detection was performed using the HaloPlex HS target enrichment system (Agilent Technologies). Custom probe panels for Illumina sequencers were designed using the SureDesign online tool (Agilent Technologies). The libraries were sequenced with 150-bp paired reads using a MiSeq sequencer (Illumina) and MiSeq Reagent Kit v2 (300 cycles) (Illumina). Obtained fastq files were analysed by Agilent SureCall 3.5 (Agilent) using the default settings. In brief, cleaned fastq files were mapped onto the human reference genome version GRCh37 (hg19) with BWA-MEM. SNVs were called using the SNPPET SNP caller. High-confidence regions with at least 250 total reads and 5 or more reads with SNVs were used for analysis. Raw variants were annotated by SureCall with default settings. After the SureCall analysis, SnpEff (v.4.3t) annotation was performed to specifically detect splicing donors and acceptors. Raw variants were filtered by removing variants registered in HGVD and common variants detected in the pooled targeted sequencing data of four blood samples in this study. Hotspot mutations were defined as SNVs with more than 30 registered cases in the COSMIC database. D437, S438 and R512* mutations in *ZC3H12A* were also included as hotspot mutations. To exclude potential sequencing errors, we selected truncation mutations that were uniquely identified in our targeted sequencing study except for hotspot mutations. These data were also used to calculate dN/dS ratios using the R package dndscv (v.0.0.1.0). Mutations detected by target sequencing and the related data are available in Supplementary Table 2.

Quantitative PCR with reverse transcription

To analyse the induction of *NFKBIZ* and *NOS2* by treatment with IL-17A, we treated patient-derived wild-type, *NFKBIZ*-mutant, *TRAF3IP*-mutant and *ZC3H12A*^{D437Y} organoids, as well as genetically engineered *NFKBIZ*^{KO}, *TRAF3IP2*^{KO}, *IL17RA*^{KO} and *ZC3H12A*$^{D437Y\text{-}KI}$ organoids, with 100 ng ml^{-1} IL-17A (R&D systems) for 3 h(for *NFKBIZ*) and 24 h (for *NOS2*). In specific experiments, the organoids were expanded in a Noggin-deficient condition. Total RNA was extracted from organoids using the RNeasy Mini Kit (Qiagen) and reverse-transcribed using the Omniscript RT Kit (Qiagen) according to the manufacturer's instructions. Quantitative PCR was performed using FastStart Essential DNA Probes Master (Roche) and Universal Probe Library (UPL) probes, and fluorescence was detected with a LightCycler 96 device (Roche). Relative gene expression levels were calculated with the $\Delta\Delta C_T$ method using *ACTB* as a housekeeping gene. UPL probe numbers and primer sequences are available in Supplementary Table 3.

Capillary-based immunodetection

Patient-derived *NFKBIZ*-mutant (UC02N_IS), *TRAF3IP2*-mutant (UC07T_IT) and *ZC3H12A*^{D437Y} (UC23N_IS) organoids, as well as genetically engineered *NFKBIZ*^{KO} organoids, together with their wild-type counterparts (UC02N_UA, UC07T_IR and UC23N_UA), were treated with or without 100 ng ml^{-1} IL-17A for 24 h and collected using Cell Recovery Solution (Corning). Total protein was extracted using Cell Lysis Buffer (Cell Signaling Technology) supplemented with Halt Protease and Phosphatase Inhibitor Cocktail (Thermo Fisher Scientific). IκBζ, PIGR, IL-17RA and β-actin proteins were detected using a Wes capillary-based device (Protein Simple) according to the manufacturer's instructions. Rabbit anti-IκB-ζ (9244, Cell Signaling Technology), rabbit anti-PIGR (HPA012012, Sigma-Aldrich), mouse anti-IL-17RA (sc-376374, Santa Cruz), mouse anti-β-actin (A1978, Sigma-Aldrich) antibodies were used with a 1:50 dilution ratio. Signals were detected by HRP-conjugated secondary goat anti-rabbit (042-206, Protein Simple) and goat anti-mouse (042-205, Protein Simple) antibodies.

Immunohistochemistry and immunocytochemistry

Sections of 5-μm thickness were prepared from archived paraffin-embedded tissues. Immunostaining of PIGR, secretory IgA, cytokeratin 20 and iNOS was performed following standard procedures using rabbit anti-PIGR (HPA012012, Sigma-Aldrich, 1:3,500), mouse anti-IgA secretory component (ab3924, Abcam, 1:1,200), mouse anti-cytokeratin 20 (M7019, DAKO, 1:100), and rabbit anti-iNOS (ab15323, Abcam, 1:100) antibodies. Whole-mount immunofluorescent staining of organoids was performed as previously described[40]. For immunostaining, rabbit anti-cleaved caspase-3 (Asp175) (9661, Cell Signaling Technology, 1:400) antibody was used, with subsequent labelling by Alexa Fluor 568-conjugated anti-rabbit antibody (Thermo Fisher Scientific). Nuclei were counterstained with Hoechst 33342 (Thermo Fisher Scientific).

CRISPR-based knockout screening

We selected *NFKIBZ*, *IL17RA* and *PIGR* genes for knockout screening. Gene targets and guide RNA sequences are listed in Supplementary Table 3. For each gene, four or five sgRNA targets were designed, and ten negative control sequences without homologous sequences in the human genome were also included. The lentiGuide-Puro vector (Addgene, 52963) was modified to increase the knockout efficiency, as previously described[46,47]. Custom sgRNA oligonucleotides were pooled and cloned into the modified vector. Lentiviral particles that express human codon-optimized SpCas9 or pooled sgRNAs were produced by a transient transfection of the lentiCas9-Blast (Addgene, 52962) or modified lentiGuide-Puro vector. We used two normal colonic organoid lines that we previously established as normal counterparts of colon

cancer organoids[38]. The organoids were first infected with the Cas9-expressing lentivirus by spinoculation[48]. To retain Cas9 expression, 10 μg ml^{-1} blasticidin S (FUJIFILM Wako Pure Chemical Corporation) was added to the medium throughout the experiment. The organoids were subsequently infected with the pooled sgRNA lentivirus via spinoculation at a multiplicity of infection of 0.2 and then selected with 2 μg ml^{-1} puromycin for two days. The infected organoids were allowed to expand for four weeks either in the presence or absence of 100 ng ml^{-1} IL-17A with weekly passaging using TrypLE Express. After the treatment, DNA was collected using the QIAamp Blood Mini Kit (Qiagen), and the sequencing library was prepared as previously described with two steps of PCR amplification[49]. The amplicons were sequenced with MiSeq (Illumina) and the reads without any matching sgRNA sequence were filtered out. The abundance of each sgRNA was calculated as the proportion of the sgRNA read count to the total filtered reads. The fold-change value for each sgRNA was defined as the ratio of the sgRNA frequency in IL-17A-treated organoids to that in IL-17A-untreated organoids, and the values were further normalized to the mean fold-change value of non-targeting controls.

Genetic engineering of organoids

Knockout of the *NFKBIZ*, *IL17RA* and *TRAF3IP2* genes was performed using electroporation as previously described[50] with slight modifications. Guide RNA sequences are listed in Supplementary Table 3. The organoids were cultured with the modified culture protocol using IGF-1 and FGF-2[37] throughout the experiment. Each sgRNA was co-delivered with the GFP-puro piggyBAC transposon vector (PB513B-1, System Biosciences). We also adopted a two-day hypothermic incubation at 30 °C to facilitate genome editing[40,51]. Five days after electroporation, organoids were treated with 2 μg ml^{-1} puromycin (Thermo Fisher Scientific) for two days for selection. Puromycin-resistant clones were manually picked up, expanded and subjected to sequence confirmation by bacterial cloning or sequence trace decomposition[52]. Loss of the IκBζ protein was confirmed for all knockout lines using capillary-based immunodetection.

Knock-in of *ZC3H12A* mutations was done by homology-directed repair using sgRNAs and targeting vectors. The 5′ and 3′ arms that flank the DSGXXS motif were designed independently owing to the high GC content in the 3′ arm. The 5′ homology arm with a D437Y mutation (GAC>TAC) (1 kb in length) was synthesized with GeneArt Strings DNA Fragments (Thermo Fisher Scientific). The 3′ arm (1 kb in length) was amplified by genomic PCR. The arms were then cloned into the pCR2.1-TOPO vector (Thermo Fisher Scientific) using the In-Fusion HD cloning kit (Clonetech). sgRNAs and the targeting vector were co-electroporated with the GFP-puro piggyBAC transposon vector in a similar manner to the knockout experiment. After puromycin selection, single organoid clones were manually isolated, expanded and subjected to sequencing. Owing to the low knock-in efficiency, all the clones with a knocked-in allele contained frameshift mutations in the other allele. We therefore used a clone heterozygous for a knocked-in allele and an in-frame frameshift mutation for analysis.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Gene expression datasets are available from the Gene Expression Omnibus (GEO) under the accession code GSE127757. Whole-exome sequencing data have been deposited to the Japanese Genotype-phenotype Archive under the accession number JGAS00000000199. Source data for Figs. 1, 3 and 4 and Extended Data Figs. 4, 6, 8 and 9 are provided with the paper. All relevant data that are included with this study are available from the corresponding author on reasonable request.

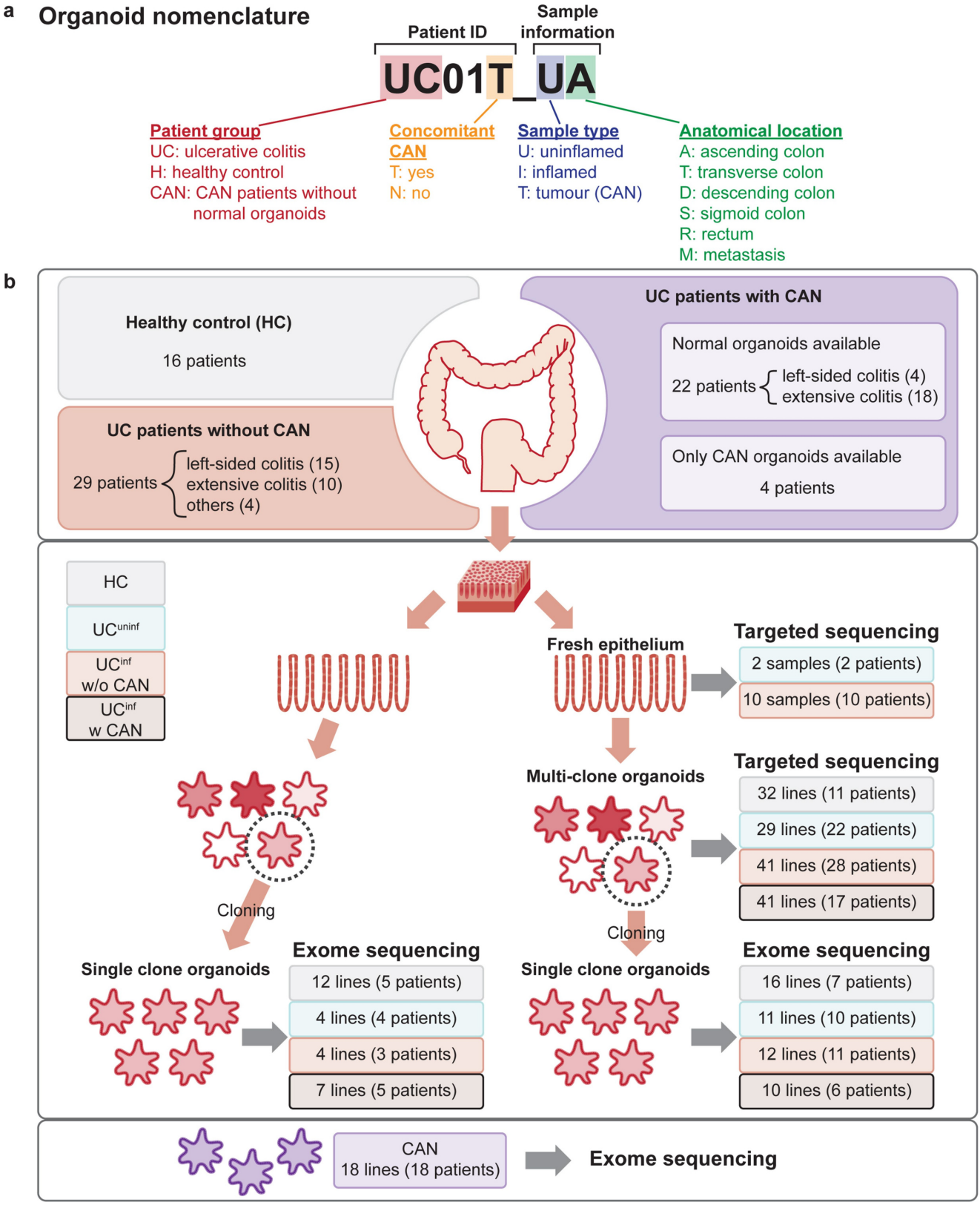
36. Magro, F. et al. Third European evidence-based consensus on diagnosis and management of ulcerative colitis. Part 1: definitions, diagnosis, extra-intestinal manifestations, pregnancy, cancer surveillance, surgery, and ileo-anal pouch disorders. *J. Crohns Colitis* **11**, 649–670 (2017).
37. Fujii, M. et al. Human intestinal organoids maintain self-renewal capacity and cellular diversity in niche-inspired culture condition. *Cell Stem Cell* **23**, 787–793 (2018).
38. Fujii, M. et al. A colorectal tumor organoid library demonstrates progressive loss of niche factor requirements during tumorigenesis. *Cell Stem Cell* **18**, 827–838 (2016).
39. Mihara, E. et al. Active and water-soluble form of lipidated Wnt protein is maintained by a serum glycoprotein afamin/α-albumin. *eLife* **5**, e11621 (2016).
40. Seino, T. et al. Human pancreatic tumor organoids reveal loss of stem cell niche factor dependence during disease progression. *Cell Stem Cell* **22**, 454–467 (2018).
41. Drost, J. et al. Sequential cancer mutations in cultured human intestinal stem cells. *Nature* **521**, 43–47 (2015).
42. Smyth, G. K. Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 1544-6115 (2004).
43. Planell, N. et al. Transcriptional analysis of the intestinal mucosa of patients with ulcerative colitis in remission reveals lasting epithelial cell alterations. *Gut* **62**, 967–976 (2013).
44. Vanhove, W. et al. Strong upregulation of AIM2 and IFI16 inflammasomes in the mucosa of patients with active inflammatory bowel disease. *Inflamm. Bowel Dis.* **21**, 2673–2682 (2015).
45. Ramilowski, J. A. et al. A draft network of ligand-receptor-mediated multicellular signalling in human. *Nat. Commun.* **6**, 7866 (2015).
46. Chen, B. et al. Dynamic imaging of genomic loci in living human cells by an optimized CRISPR/Cas system. *Cell* **155**, 1479–1491 (2013).
47. Tzelepis, K. et al. A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. *Cell Rep.* **17**, 1193–1205 (2016).
48. Koo, B. K. et al. Controlled gene expression in primary *Lgr5* organoid cultures. *Nat Methods* **9**, 81–83 (2012).
49. Shalem, O. et al. Genome-scale CRISPR–Cas9 knockout screening in human cells. *Science* **343**, 84–87 (2014).
50. Fujii, M., Matano, M., Nanki, K. & Sato, T. Efficient genetic engineering of human intestinal organoids using electroporation. *Nat. Protocols* **10**, 1474–1485 (2015).
51. Guo, Q. et al. 'Cold shock' increases the frequency of homology directed repair gene editing in induced pluripotent stem cells. *Sci. Rep.* **8**, 2080 (2018).
52. Hsiau, T. et al. Inference of CRISPR edits from Sanger trace data. Preprint at *bioRxiv* https://doi.org/10.1101/251082 (2019).

Acknowledgements This work was supported by AMED (grant numbers JP19cm0106206 and JP19gm5010002), AMED-CREST (grant number JP19gm1210001) and the JSPS KAKENHI (grant numbers JP17H06176 and JP26115007). K.T., Y.O. and M.F. were supported by the Japan Society for the Promotion of Science Research Fellowships for Young Scientists. We also thank the Collaborative Research Resources at the School of Medicine, Keio University for providing technical assistance.

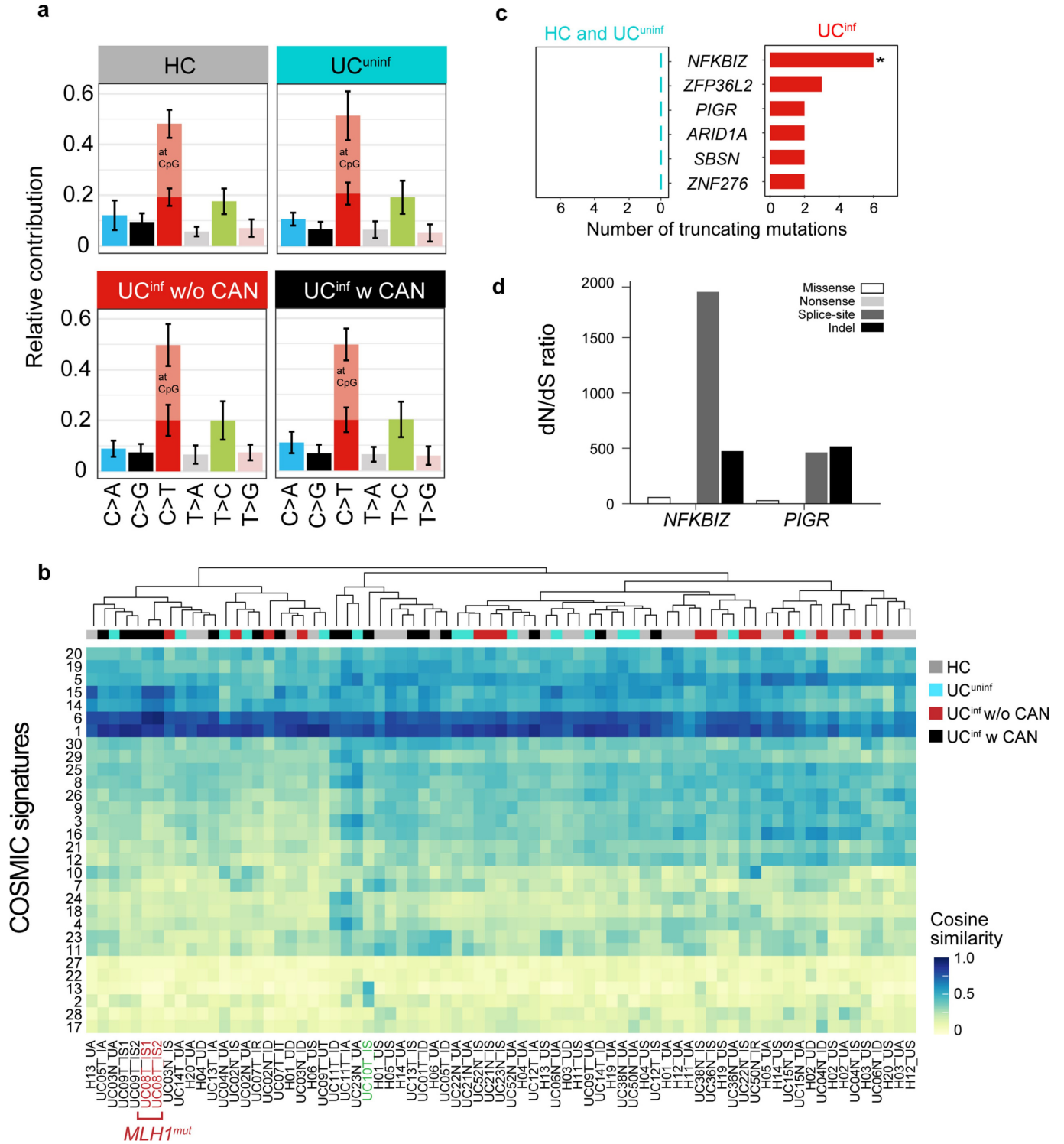
Author contributions K.N., M.F. and T.S. designed the study. T.H., S.I. and T.K. provided specimens. Y.O. and S.T. performed histochemical analysis. K.N., M.F., M.M., S.S., K.K. and K.I. performed sample preparation. K.N., M.F., M.M., S.N., A.T., S.S., K.I., K.Y. and N.S. performed organoid experiments. M.F., M.S., S.D., Y.N. and K.T. performed sequencing and bioinformatics analysis. R.I. supervised statistical analyses. K.N., M.F. and T.S. wrote the manuscript.

Competing interests T.S. is an inventor on several patents related to organoid culture.

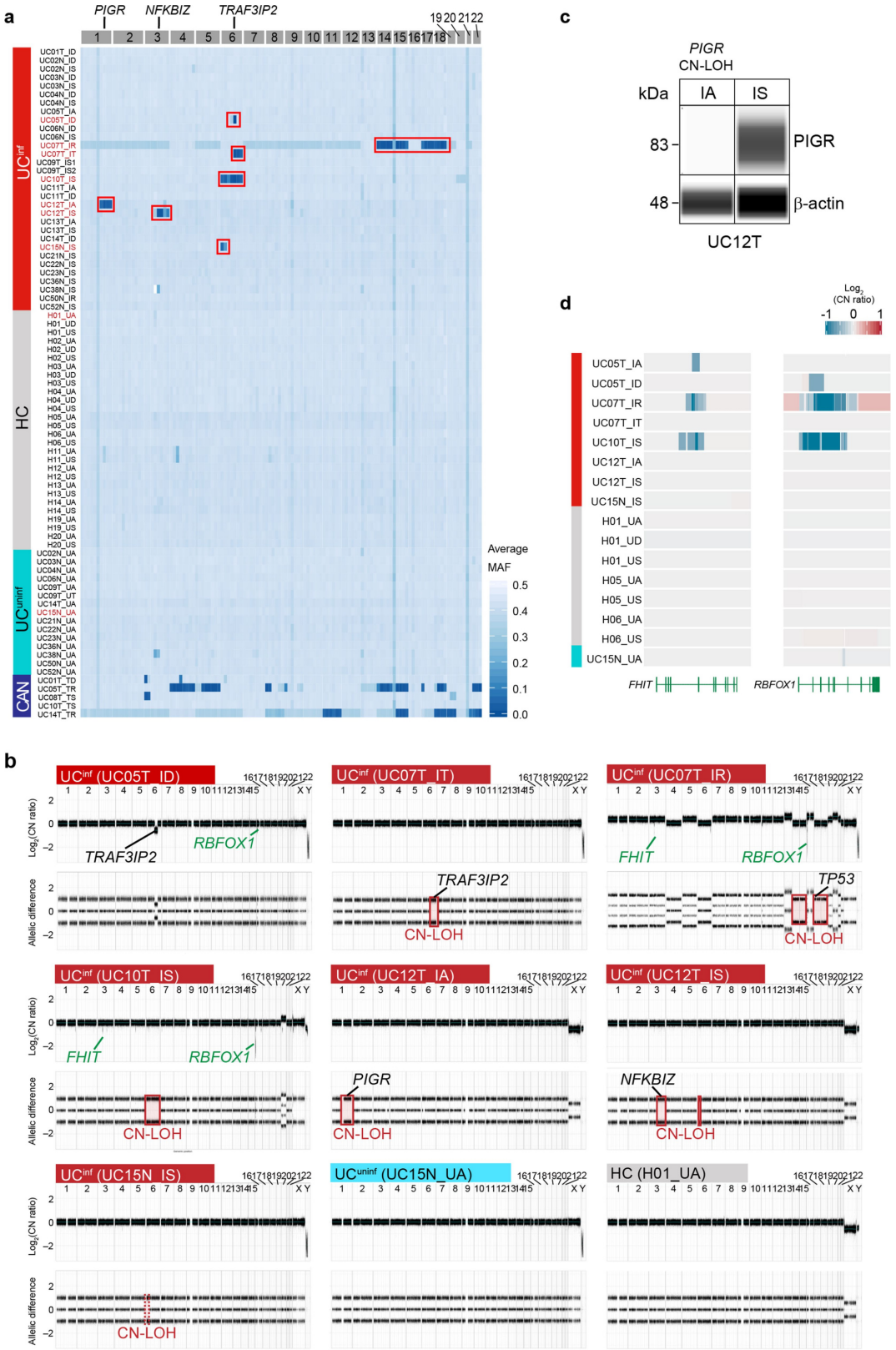
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1844-5.
Correspondence and requests for materials should be addressed to T.S.
Peer review information *Nature* thanks Judy Cho, Ramnik Xavier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



Extended Data Fig. 1 | Study design. a, Nomenclature of organoids. **b**, Patient demographics and the breakdown of the organoids used for each analysis.

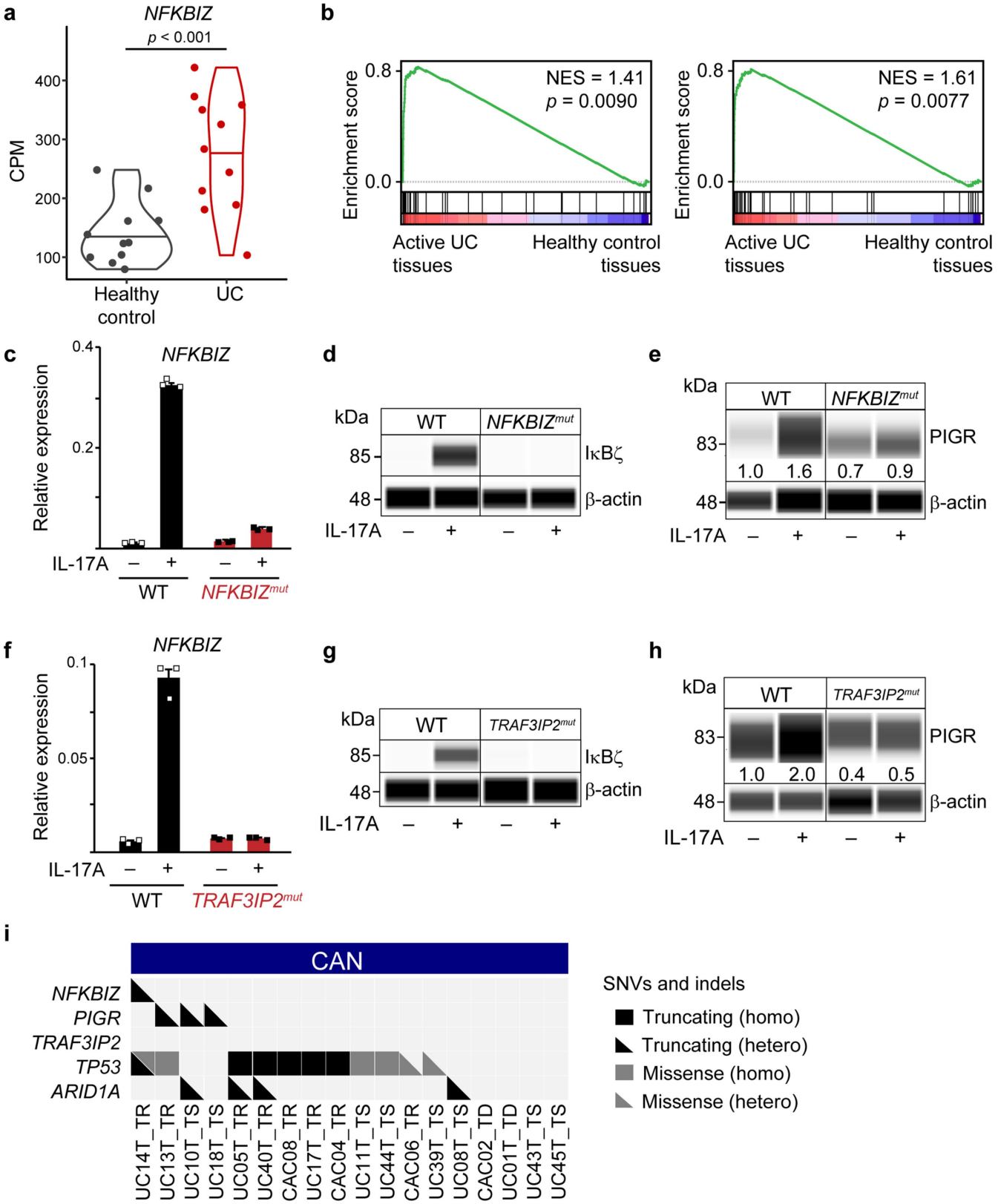


Extended Data Fig. 2 | Mutational signatures of HC, UC and CAN organoids. **a**, Relative contribution of the nucleotide transition patterns in each sample group in healthy ($n = 28$), UC^{uninf} ($n = 15$), UC^{inf} without CAN ($n = 15$) and UC^{inf} with CAN ($n = 17$) organoids. Data are shown as the mean (±s.d.) relative contribution of base substitution types in each sample type. **b**, Cosine similarity of the mutational signature of each organoid clone to the COSMIC signatures (1–30). UC^{inf} organoids with an *MLH1* mutation (UC08T_IS1, UC08T_IS2) are highlighted in red. One organoid line with a contribution of APOBEC signatures (signatures 2 and 13) is highlighted in green characters (UC10T_IS). **c**, Number of truncating mutations in the indicated genes in healthy, UC^{uninf} ($n = 43$) and UC^{inf} ($n = 31$) organoids. Genes with two or more organoid clones that have truncating mutations are shown. *MLH1*-mutated organoids (UC08T_IS1, 2) were excluded from this analysis. *$P = 0.004$; $q < 0.1$ (two-sided Fisher's exact test with Benjamini–Hochberg adjustment). **d**, dN/dS ratios of *NFKBIZ* and *PIGR* (genes with pglobal < 0.05 and qglobal < 0.1), calculated using the whole-exome sequencing data.



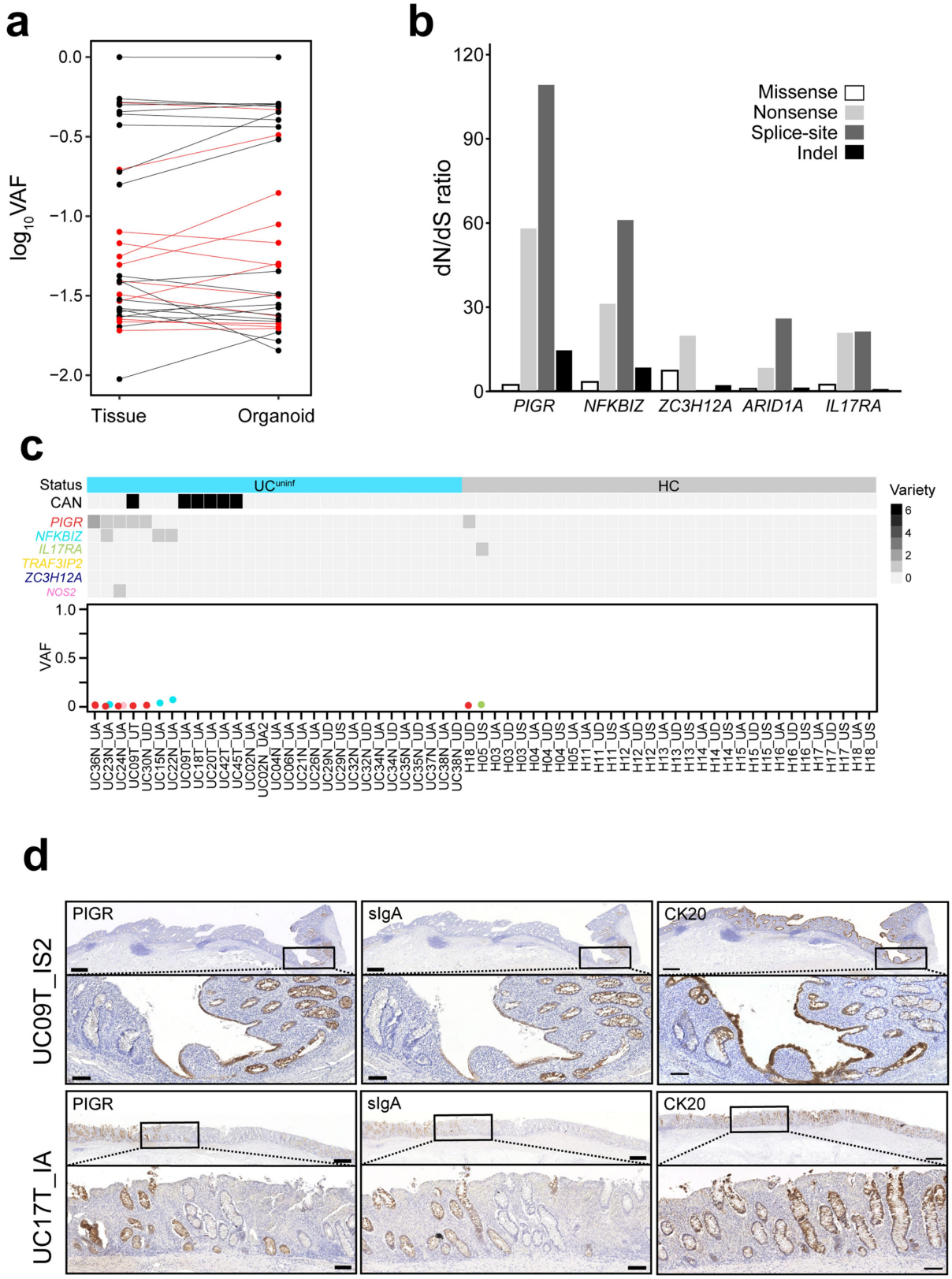
Extended Data Fig. 3 | See next page for caption.

Extended Data Fig. 3 | Copy-number analysis of UC organoids. a, Heat map showing the average minor allele frequencies (MAFs) of SNPs that were heterozygous in the normal counterpart (bin size 25 Mb). Genomic regions with LOH are indicated with red boxes. Genomic positions of *PIGR*, *NFKBIZ* and *TRAF3IP2* are shown at the top. Organoid lines that were subjected to CGH–SNP array analysis are highlighted in red. Organoid lines with *MLH1* mutations (UC08) are not shown. **b**, DNA copy-number analysis using CGH–SNP microarray analysis. For each organoid line, the genome-wide chromosome state is demonstrated by the $\log_2$-transformed signal ratio (top) and the allelic difference (bottom) values. Genomic regions with CN-LOH are shown in red. Deletions in common fragile sites are shown in green. Genetic lesions associated with UC-related mutations are shown in black. **c**, Loss of PIGR protein in UC12T_IA organoids with CN-LOH at the *PIGR* locus. Intact PIGR protein expression in UC12T_IS organoids is also shown. β-actin was used as a loading control. Data are representative of $n = 2$ technical replicates with similar results. **d**, Focal deletion of fragile sites in UC^{inf} organoids, detected by CGH analysis. Chromosome states are visualized using the $\log_2$-transformed signal ratio. Deletions in chromosomal fragile sites at positions 3p14 and 16p13.2 were uniquely detected in the UC^{inf} epithelium.



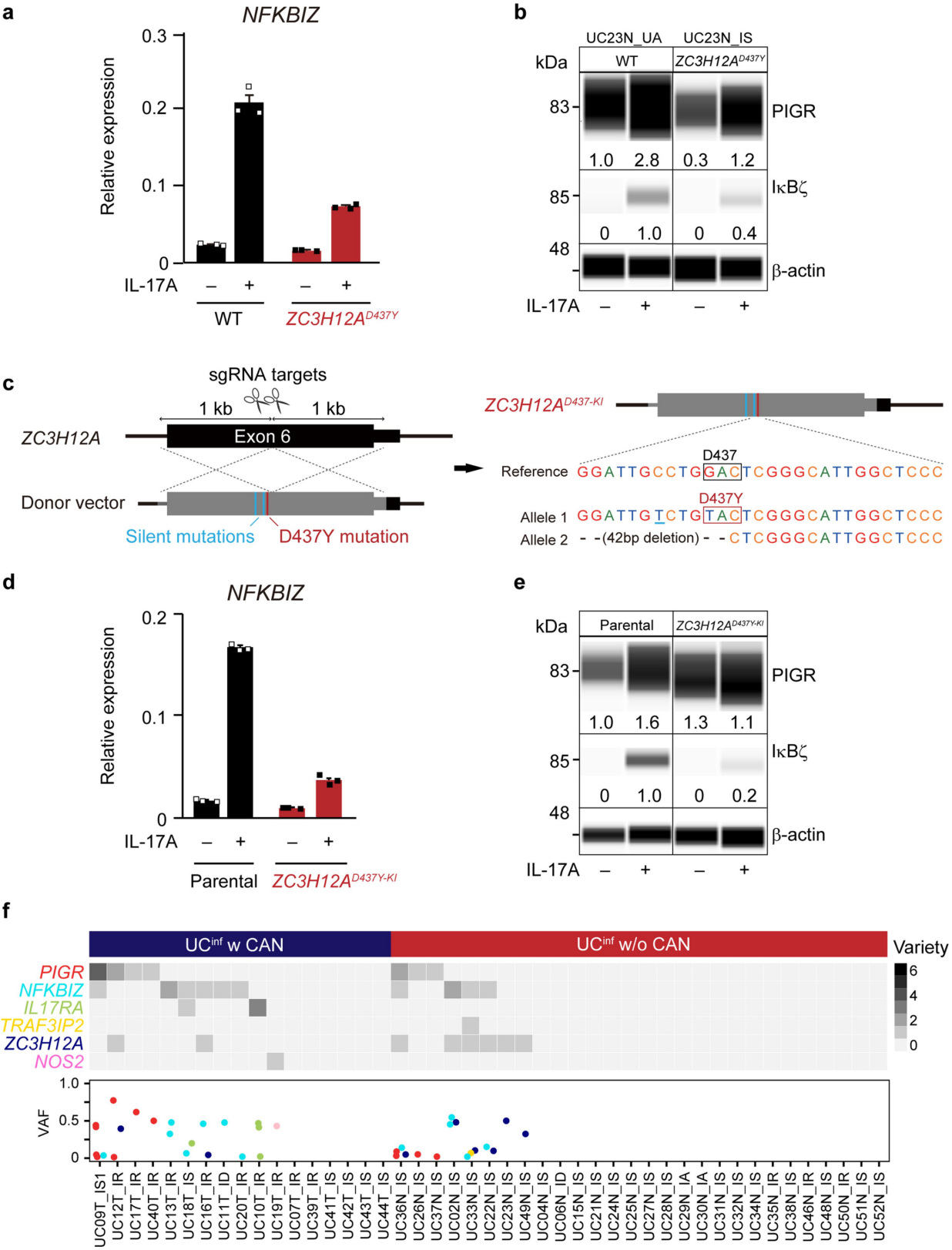
Extended Data Fig. 4 | See next page for caption.

Extended Data Fig. 4 | Characterization of UC-related mutations. a, Increased expression of *NFKBIZ* mRNA in purified UC epithelium (E-MTAB-5464). *NFKBIZ* expression was compared between healthy control individuals ($n = 11$) and patients with UC ($n = 11$) using the exact negative binomial test in edgeR ($P = 5.0 \times 10^{-6}$). Bars represent the median. **b**, Gene set enrichment analysis using public datasets (GSE38713, left; GSE59071, right) and in-house IL-17 target genes (Supplementary Table 3). NES; normalized enrichment score. Left, $n = 15$ (active colitis), $n = 13$ (healthy control); right, $n = 82$ (active colitis), $n = 11$ (healthy control). **c**–**h**, IL-17A-induced expression of *NFKBIZ* (IκBζ) and PIGR in colon organoids. Organoids were stimulated with 100 ng ml^{-1} IL-17A for 3 h (**c**, **f**) or 24 h (**d**, **e**, **g**, **h**). The expression of *NFKBIZ* (IκBζ) and PIGR in control, *NFKBIZ*-mutant and *TRAF3IP2*-mutant organoids was analysed by quantitative PCR with reverse transcription (**c**, **f**) and immunoassay (**d**, **e**, **g**, **h**). Matched UC^{uninf} organoids that were derived from the same patients were used as controls. mRNA expression levels are shown relative to the expression of *ACTB*. Expression levels of PIGR protein relative to β-actin are shown at the bottom of the pseudo-blot images. Results are shown from two (**c**–**e**) and one (**f**–**h**) biologically independent experiments, and two technical replicates (**g**, **h**). Data are shown as mean + s.e.m. (**c**, **f**). **i**, Genetic mutations detected by whole-exome sequencing in CAN organoids. Patterns of mutations are specified using the same symbols as Fig. 2a.



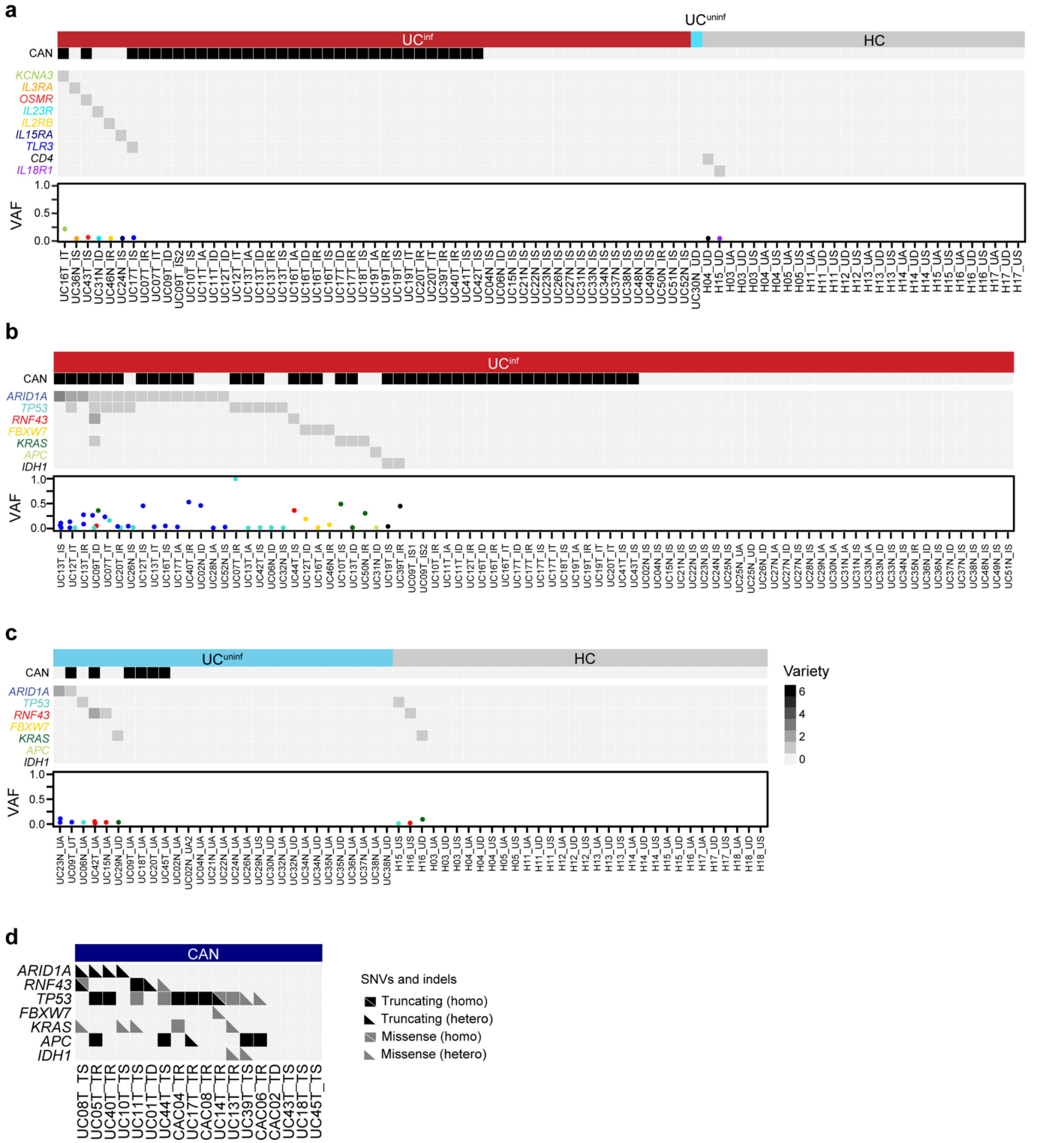
Extended Data Fig. 5 | See next page for caption.

Extended Data Fig. 5 | Targeted sequencing of UC-related genes. a, Consistency of VAFs between fresh epithelium samples and organoids ($n = 11$ pairs). Red dots and lines show truncating and non-synonymous mutations in UC-mutated genes (*NFKBIZ*, *PIGR*, *IL17RA*, *TRAF3IP2* and *ZC3H12A*) ($n = 12$ mutations). Black dots and lines show synonymous mutations in genes that are mutated in UC and any mutations in the other genes ($n = 19$ mutations). VAFs did not significantly differ between the sample pairs ($P = 1$, two-sided Wilcoxon signed-rank test). **b**, dN/dS ratios of the indicated genes in the targeted sequencing data. Genes with with q-values <0.1 (that is, statistically significant) are shown. **c**, Genetic mutations in UC^{uninf} and control organoids, detected by targeted sequencing. Top, for each gene, the number of different variants is shown by density. Bottom, the VAF in each analysed sample. Each dot represents the fraction of each altered allele per total reads mapped to the gene. Genes are shown in different colours as specified on the left. **d**, PIGR (top), secretory IgA (middle) and cytokeratin20 (CK20) (bottom) immunostaining of UC^{inf} epithelium (UC09T_IS2 and UC17T_IA) with focal loss of PIGR and secretory IgA expression. Magnified areas show boundaries between regions in which PIGR expression is intact and regions in which PIGR expression is lost. Images are representative images of four regions in which PIGR expression is lost, in eight patients with UC. Scale bars, 500 µm (low magnification); 100 µm (high magnification).

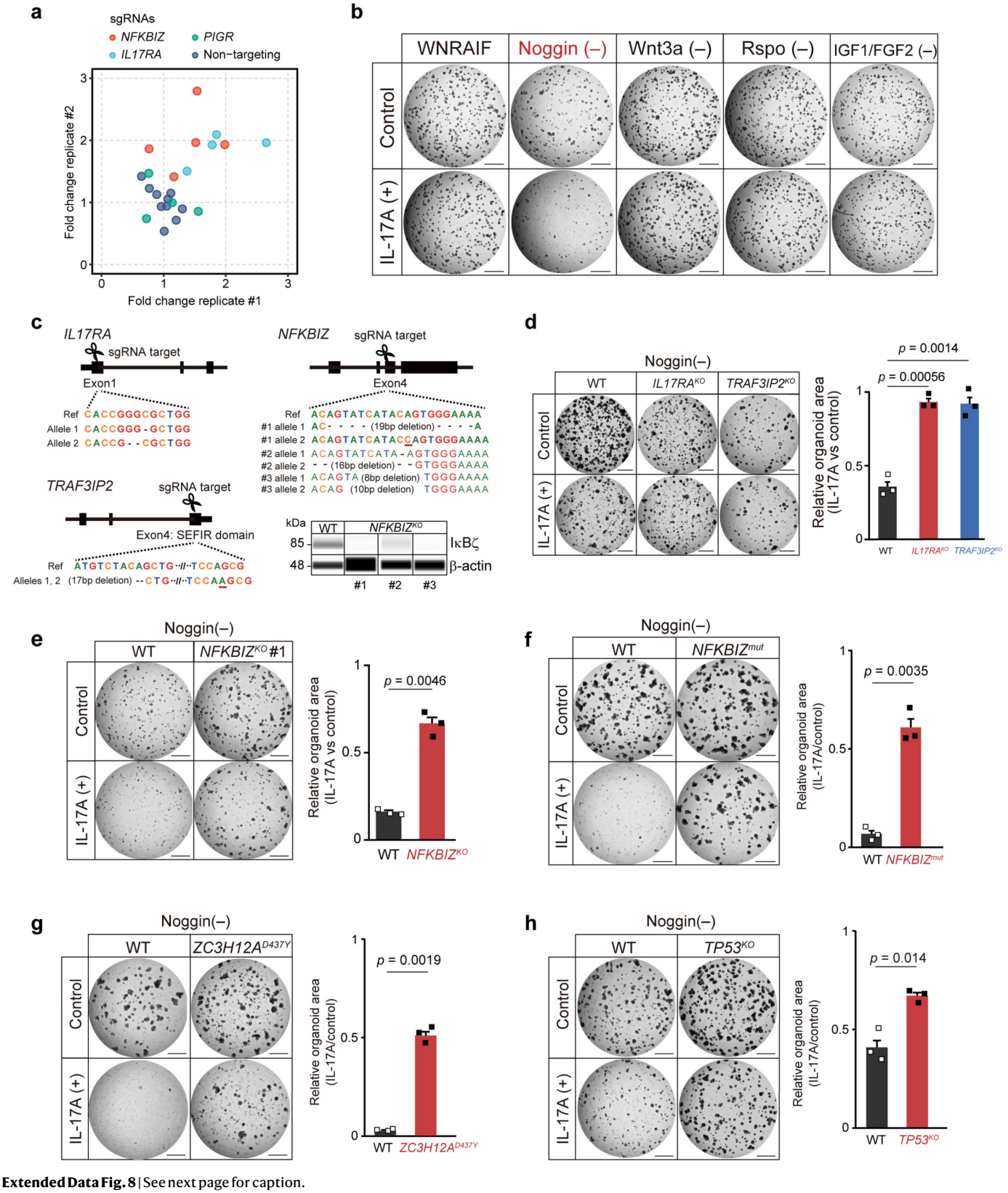


Extended Data Fig. 6 | See next page for caption.

Extended Data Fig. 6 | Targeted sequencing of UC-related genes (continued). **a**, **b**, Reduced response to treatment with IL-17A in patient-derived cloned *ZC3H12A*^{D437Y} UC^{inf} organoids (UC23N_IS) in comparison to wild-type counterpart organoids (UC23N_UA). Results are representative of $n = 2$ technical replicates with similar results. **c**, Design of CRISPR–Cas9-mediated *ZC3H12A*^{D437Y} knock-in. The donor vector contains two silent mutations at the protospacer adjacent motifs flanking the CRISPR targets, in addition to the D437Y mutation. Each allele of the knocked-in clone used for analysis is shown on the right. The clone is heterozygous for a knocked-in D437Y mutation and a 42-bp in-frame deletion. **d**, **e**, Reduced response to treatment with IL-17A in genetically engineered *ZC3H12A*$^{D437Y\text{-}KI}$ organoids compared to parental wild-type organoids. Results are representative of $n = 2$ technical replicates with similar results. **f**, Per-patient data of the targeted sequencing result. For each patient, mutations in organoids that were derived from the most-distal sample are shown. Top, for each gene, the number of different variants is shown by density. Bottom, dots represent the VAF of each variant. Genes are shown in different colours as specified on the left. Data are shown as mean + s.e.m. (**a**, **d**).

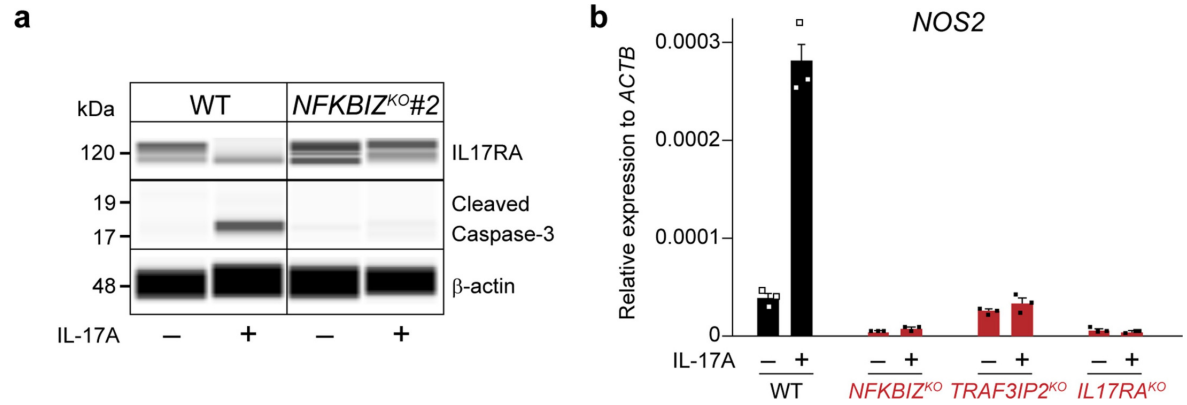


Extended Data Fig. 7 | Targeted sequencing of cytokine and PAMP receptor genes and cancer driver genes. a–**c**, Mutations in cytokine and PAMP receptor genes (**a**) and sporadic colon cancer driver genes (**b**, **c**) identified in UC^{inf}, UC^{uninf} and control organoids by targeted sequencing. Number of different variants per gene are shown by density (top). Top, for each gene, the number of different variants is shown by density. Bottom, dots represent the VAF of each variant. Genes are shown in different colours as specified on the left. **d**, Sporadic colon cancer gene mutations identified in CAN organoids.



Extended Data Fig. 8 | See next page for caption.

Extended Data Fig. 8 | IL-17A-induced cytotoxic response in human colon organoids. a, Fold-change values of individual sgRNAs in two experiments that used different lines of organoids. sgRNAs are coloured according to their targeting genes, as specified at the top. **b**, Response of wild-type organoids to treatment with IL-17A in differing growth factor conditions. The organoids showed a cytotoxic response to treatment with IL-17A only in the absence of Noggin. The WNRAIF condition is the medium condition including WNT3A, Noggin, R-spondin-1 (Rspo), A83-01, IGF-1 and FGF-2, but lacking EGF. Data are representative of $n = 2$ biologically independent experiments with similar results. **c**, sgRNA targets and confirmation of CRISPR–Cas9-mediated knockout of *IL17RA*, *TRAF3IP2* and *NFKBIZ*. For knockout of *NFKBIZ*, three *NFKBIZ*^{KO} lines (#1–3) were generated from different parental organoids. Successful knockout was confirmed by Sanger sequencing in all lines. Loss of IκBζ protein was also confirmed by an immunoassay using an organoid stimulated with IL-17A for 24 h. Result is from a single experiment. **d**, *IL17RA*^{KO} and *TRAF3IP2*^{KO} organoids are resistant to treatment with IL-17A in a Noggin-deficient condition. Right, dots show the relative area of organoids in each well. Results are representative of $n = 3$ technical replicates. **e**, *NFKBIZ*^{KO} organoid line 1 tolerates treatment with IL-17A. The other two knockout lines showed a similar response. **f**, Patient-derived *NFKBIZ*-mutant organoids (UC02N_IS) are resistant to treatment with IL-17A. Wild-type counterpart organoids (UC02N_UA) were used as a control. Data are representative of $n = 2$ biologically independent experiments with similar results. **g**, Patient-derived *ZC3H12A*^{D437Y}-mutant organoids (UC23N_IS) are tolerant of treatment with IL-17A compared to their wild-type counterparts (UC23N_UA). Data are representative of $n = 2$ technical replicates with similar results. **h**, Knockout of *TP53* mitigates IL-17A-induced cytotoxicity. Data are representative of $n = 2$ biologically independent experiments with similar results. Mean + s.e.m (**d**–**h**). *P* values determined by two-sided Welch's *t*-test. Scale bars, 1 mm (**b**, **d**–**h**).



Extended Data Fig. 9 | iNOS-dependent apoptosis in human colon organoids treated with IL-17A. a, Detection of cleaved caspase-3 by immunoassay. Treatment with IL-17A increases the expression of cleaved caspase-3 in wild-type, but not *NFKBIZ*^{KO}, organoids in the absence of Noggin. **b**, Analysis of *NOS2* transcript expression by quantitative PCR. IL-17A induces expression of *NOS2* transcript in wild-type, but not in *NFKBIZ*^{KO}, *TRAF3IP2*^{KO} and *IL17RA*^{KO}, organoids in the absence of Noggin.

Extended Data Table 1 | Clinical information on the participants in the study

ID	Disease duration	Age	Sex	UC extent or patient status	Sample collection	Counterpart Normal
UC01T	25	62	F	Extensive	Surgery	Blood
UC02N	13	54	F	Left sided	Biopsy	Blood
UC03N	6	61	F	Left sided	Biopsy	Blood
UC04N	32	45	M	Left sided	Biopsy	Blood
UC05T	12	48	F	Extensive	Biopsy	Blood
UC06N	27	43	M	Left sided	Biopsy	Blood
UC07T	12	37	F	Extensive	Biopsy	Blood
UC08T	9	62	M	Extensive	Surgery	Blood
UC09T	12	43	F	Left sided	Surgery	Blood
UC10T	40	62	M	Extensive	Surgery	Blood
UC11T	35	76	F	Extensive	Surgery	Blood
UC12T	16	61	M	Extensive	Surgery	Blood
UC13T	16	30	M	Extensive	Surgery	Blood
UC14T	10	74	M	Left sided	Surgery	Blood
UC15N	6	38	F	Left sided	Biopsy	Blood
UC16T	31	57	F	Extensive	Surgery	None
UC17T	12	51	F	Extensive	Surgery	None
UC18T	19	37	M	Extensive	Biopsy	None
UC19T	16	52	M	Extensive	Surgery	None
UC20T	14	55	F	Extensive	Surgery	None
UC21N	12	36	F	Left sided	Biopsy	Blood
UC22N	24	48	M	Left sided	Biopsy	Blood
UC23N	18	38	F	Left sided	Biopsy	Blood
UC24N	23	46	F	Extensive	Biopsy	None
UC25N	6	23	M	Extensive	Biopsy	None
UC26N	15	47	F	Left sided	Biopsy	None
UC27N	9	35	M	Extensive	Biopsy	None
UC28N	2	39	F	Extensive	Biopsy	None
UC29N	3	56	F	Right sided +proctitis	Biopsy	None
UC30N	10	38	F	Right sided +proctitis	Biopsy	None
UC31N	41	67	M	Extensive	Biopsy	None
UC32N	6	56	M	Left sided	Biopsy	None
UC33N	14	65	M	Extensive	Biopsy	None
UC34N	8	75	M	Left sided	Biopsy	None
UC35N	1	20	M	Proctitis	Biopsy	None
UC36N	16	29	M	Left sided	Biopsy	Blood

ID	Disease duration	Age	Sex	UC extent or patient status	Sample collection	Counterpart Normal
UC37N	5	29	M	Extensive	Biopsy	None
UC38N	27	58	M	Left sided	Biopsy	Blood
UC39T	17	35	M	Extensive	Biopsy	None
UC40T	18	34	F	Extensive	Biopsy	None
UC41T	11	68	M	Extensive	Biopsy	None
UC42T	10	70	M	Left sided	Biopsy	None
UC43T	18	50	M	Extensive	Biopsy	None
UC44T	1	68	M	Extensive	Surgery	None
UC45T	22	47	M	Left sided	Surgery	None
UC46N	22	71	F	Left sided	Biopsy	None
UC48N	25	53	M	Extensive	Biopsy	None
UC49N	9	60	M	Extensive	Biopsy	None
UC50N	22	50	M	Proctitis	Biopsy	Blood
UC51N	15	50	M	Extensive	Biopsy	None
UC52N	15	34	F	Left sided	Biopsy	Blood
H01	NA	57	M	Healthy	Biopsy	Blood
H02	NA	48	M	Healthy	Biopsy	Blood
H03	NA	44	F	Healthy	Biopsy	Blood
H04	NA	50	M	Healthy	Biopsy	Blood
H05	NA	65	F	Healthy	Biopsy	Blood
H06	NA	75	M	Healthy	Biopsy	Blood
H11	NA	67	M	Healthy	Biopsy	Blood
H12	NA	58	M	Healthy	Biopsy	Blood
H13	NA	57	F	Healthy	Biopsy	Blood
H14	NA	77	M	Healthy	Biopsy	Blood
H15	NA	70	F	Healthy	Biopsy	None
H16	NA	54	F	Healthy	Biopsy	None
H17	NA	70	F	Healthy	Biopsy	None
H18	NA	49	F	Healthy	Biopsy	None
H19	NA	32	M	Healthy	Biopsy	Blood
H20	NA	22	F	Healthy	Biopsy	Blood
CAC02	28	45	M	Extensive	Surgery	Non-tumor
CAC04	30	45	M	Left sided	Surgery	Non-tumor
CAC06	39	54	F	Extensive	Surgery	Blood
CAC08	11	47	M	Extensive	Surgery	Blood

Patient IDs and other related information in our cohort. The IDs are specified as described in Extended Data Fig. 1a.

nature research

Corresponding author(s): Toshiro Sato

Last updated by author(s): Oct 21, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	No software was used for collecting data.
Data analysis	For whole exome sequencing analysis, raw fastq reads were mapped onto the human reference genome GRCh37 (hg19) with BWA (ver 0.7.8). A GATK (ver 1.6.13) toolkit UnifiedGenotyper was used for SNV and indel calling. SnpEff (ver 3) was used for the annotation of SNPs in dbSNP build135 and the 1000 genome project database. R (ver 3.5.1) was used for HGVD (release ver 2.3.0) annotation, filtering and counting the number of mutations. lme (R package nlme, ver 3.1) was used for apply a linear mixed model to estimate the effect of age, disease duration and disease groups on the number of SNVs. MutationalPatterns (ver 1.8.0, R package) was used for mutation signature analysis. Fisher's exact test was done using the fisher.test function (R package stat, ver 3.5.1) . dndscv (ver 0.0.1.0, R package) was used for dN/dS analysis. For targeted sequencing analysis, SureCall (ver 3.5, Agilent) was used for mapping (BWA-MEM), for SNV and indel calling (SNPPET SNP Caller) and for annotation of SNPs in dbSNP and the 1000 genome project database. SnpEff (ver 4.3t) was also used for EFFECT annotation to specifically detect splicing donors and accepters. R (ver 3.5.1) was used for HGVD (release ver 2.3.0) annotation, filtering and counting the number of mutations. dndscv (ver 0.0.1.0, R package) was used for dN/dS analysis. For copy number analysis, apt-copynumber-cyto-ssa (Affymetrix Power Tools, ver 1.19.0) was used for obtaining probe-level copy number and allele difference. segmentCGH (R package rCGH, ver 1.8.1) was used for segmentation analysis. For gene expression analysis, rma (R package affy, ver 1.56.0) was used for normalization. Gene Set Enrichment Analysis module (ver 3.24.3, GenePattern) was used for GSEA. Differentially expressed genes were determined using limma (ver 3.40.2, R package). Public RNA-seq data was analyzed using edgeR (ver 3.24.3, R package).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

Gene expression datasets were deposited in Gene Expression Ominibus. Accession code of GEO is GSE127757.
The secure token of GEO is as followed :ijoxsgmqfxizbun.
Whole exome datas were deposited in JGA database. Accession number is JGAS00000000199.
All relevant data that are included with this study are available from corresponding author upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No formal statistical methods were used to predetermine sample size. Instead, we enrolled all individuals who consent to participate our study during the enrollment period between 1/Aug/2014 and 31/Mar/2019. We collected 44 and 144 colorectal samples from 16 HCs and 55 UC patients, respectively. The sample size of our study was sufficient to assess our sequencing data and to confirm novel findings with statistical significance.
Data exclusions	All the samples which we could established organoids were used for analysis. UC organoids with a MLH1 mutation were excluded from the mutation number analysis.
Replication	All attempts at replication were successful.
Randomization	Not applicable for this study because this study is a case-series study .
Blinding	The experiments and analysis used did not require blinding

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	Primary antibodies used in this study were as follows: rabbit anti-cleaved-caspase3 (Asp175) (9661, Lot: 45, Cell Signaling Technology, 1:400), rabbit anti-PIGR (HPA012012, Lot: c106465, Sigma-Aldrich, for IHC 1:3500, for capillaraly-based immunodetection 1:50), rabbit anti-IκBζ (9244, Lot: 2, Cell Signaling Technology, 1:50), mouse anti-β-Actin (AC-15) (A1978, Lot:037M4782V, Sigma-Aldrich, 1:50), mouse anti-IgA secretory component (ab3924, Lot: GR3236730-3, Abcam, 1:12000), rabbit anti-iNOS (ab15323, Lot: GR3272520-2, Abcam, 1:100), mouse anti-cytokeratin 20 (Ks20.8) (M7019, Lot: 20061091, Dako, 1:100), mouse anti-IL17R (G-9) (sc-376374, Lot: K2817, Santa Cruz, 1:50). Secondary antibody used in this study was as follows: donkey polyclonal anti-rabbit IgG (H+L), Alexa Fluor 568 (A10042, Thermo

	Fisher Scientific, 1:500), goat anti-rabbit secondary HRP conjugate (042-206, Protein Simple), goat anti-mouse secondary HRP conjugate (042-205, Protein Simple)
Validation	All antibodies used were validated by the respective vendors for the applications used in this study. The specificity of anti-PIGR, anti-IkBz and anti-igA secretory component antibodies were verified by organoid knockout experiments.

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	All organoids were established from the tissues obtained from patients at Keio University Hospital or Tokyo University Hospital.
Authentication	None of the organoids were authenticated
Mycoplasma contamination	The organoids were not tested for mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	None of the used cell lines are listed in the ICLAC database of commonly misidentified cell lines.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Participants were recruited through Keio University Hospital and Tokyo University Hospital. We enrolled a total of 16 (8 males and 8 females) HCs and 55 (33 males and 22 females) UC patients who underwent surgery and/or therapeutic or diagnostic endoscopy for lower gastrointestinal symptoms. 20-75 year-old UC patients and 22-77 year-old HCs were enrolled. Among these patients with UC, 26 had concomitant CAN and the remaining 29 were without CAN. All subjects and samples were anonymized with unique IDs before analysis. The detailed characteristics of cases are shown in Supplementary Table 3.
Recruitment	The individuals who underwent colonoscopy or surgery for gastrointestinal disease or symptoms were randomly recruited with written informed consent.
Ethics oversight	This study was approved by the ethics committees of Keio University and Tokyo University.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Article

Frequent mutations that converge on the NFKBIZ pathway in ulcerative colitis

https://doi.org/10.1038/s41586-019-1856-1

Received: 26 February 2019

Accepted: 8 November 2019

Published online: 18 December 2019

Nobuyuki Kakiuchi[1,2,3], Kenichi Yoshida[1], Motoi Uchino[4], Takako Kihara[5], Kotaro Akaki[6], Yoshikage Inoue[1,2,7], Kenji Kawada[7], Satoshi Nagayama[8], Akira Yokoyama[1,9], Shuji Yamamoto[3], Minoru Matsuura[3,10], Takahiro Horimatsu[9], Tomonori Hirano[1,2,3], Norihiro Goto[3], Yasuhide Takeuchi[1,2,11,12], Yotaro Ochi[1,2], Yusuke Shiozawa[1], Yasunori Kogure[1,13], Yosaku Watatani[1,2], Yoichi Fujii[1,2], Soo Ki Kim[1,3], Ayana Kon[1,2], Keisuke Kataoka[1,13], Tetsuichi Yoshizato[1], Masahiro M. Nakagawa[1,2], Akinori Yoda[1,2], Yasuhito Nanya[1,2], Hideki Makishima[1,2], Yuichi Shiraishi[14], Kenichi Chiba[14], Hiroko Tanaka[15], Masashi Sanada[1,16], Eiji Sugihara[17], Taka-aki Sato[17], Takashi Maruyama[18], Hiroyuki Miyoshi[19], Makoto Mark Taketo[19], Jun Oishi[20], Ryosaku Inagaki[1,20], Yutaka Ueda[20], Shinya Okamoto[21], Hideaki Okajima[21,22], Yoshiharu Sakai[7], Takaki Sakurai[11], Hironori Haga[12], Seiichi Hirota[5], Hiroki Ikeuchi[4], Hiroshi Nakase[3,23], Hiroyuki Marusawa[3], Tsutomu Chiba[3,24], Osamu Takeuchi[6], Satoru Miyano[14,15], Hiroshi Seno[3] & Seishi Ogawa[1,2,25]*

Chronic inflammation is accompanied by recurring cycles of tissue destruction and repair and is associated with an increased risk of cancer[1–3]. However, how such cycles affect the clonal composition of tissues, particularly in terms of cancer development, remains unknown. Here we show that in patients with ulcerative colitis, the inflamed intestine undergoes widespread remodelling by pervasive clones, many of which are positively selected by acquiring mutations that commonly involve the *NFKBIZ*, *TRAF3IP2*, *ZC3H12A*, *PIGR* and *HNRNPF* genes and are implicated in the downregulation of IL-17 and other pro-inflammatory signals. Mutational profiles vary substantially between colitis-associated cancer and non-dysplastic tissues in ulcerative colitis, which indicates that there are distinct mechanisms of positive selection in both tissues. In particular, mutations in *NFKBIZ* are highly prevalent in the epithelium of patients with ulcerative colitis but rarely found in both sporadic and colitis-associated cancer, indicating that *NFKBIZ*-mutant cells are selected against during colorectal carcinogenesis. In further support of this negative selection, we found that tumour formation was significantly attenuated in *Nfkbiz*-mutant mice and cell competition was compromised by disruption of *NFKBIZ* in human colorectal cancer cells. Our results highlight common and discrete mechanisms of clonal selection in inflammatory tissues, which reveal unexpected cancer vulnerabilities that could potentially be exploited for therapeutics in colorectal cancer.

Chronic inflammation is a major cause of morbidity and mortality in the human population and poses a substantial cancer risk[1,2]. A common feature of chronic inflammation is recurring tissue injury and repair, which is fuelled by long-lasting, unregulated immunity and often culminates in tissue remodelling and debilitating organ dysfunction[3]. Previous studies have shown that extensive tissue remodelling can take place in an age-dependent manner even in tissues that are apparently normal, and that this is mediated by the expansion of numerous clones that carry common cancer-related mutations[4–6]. The expansion of somatically mutated clones may also have a role in inflammation-associated remodelling of tissues, explaining the cancer risk that is associated with inflammation. However, it is largely unknown whether this actually happens—and if so, the frequency and extent of the clonal expansion, the mutations that are involved, and how it modifies the inflammatory disease process and the development of cancer.

Ulcerative colitis (UC) is a common form of inflammatory bowel disease that affects around 2.1 million people in Europe alone[7]. It is characterized by persistent inflammation of the large intestine, which causes severe destruction of the intestinal mucosa and the formation of intractable ulcers. Although the pathophysiology of UC is not fully understood, deregulated immune cells and cytokines—as well as intestinal dysbiosis—have been implicated in the persistent inflammation[8], which is associated with an increased risk (15–20%) of cancer development (colitis-associated colorectal cancer; CAC)[9]. In this study, to understand how chronic inflammation shapes tissue remodelling and how it correlates with cancer development, we investigated clonal expansion in epithelia from patients with UC and from control individuals not affected by UC, using intensive multi-regional sampling followed by unbiased detection of somatic mutations.

A list of affiliations appears at the end of the paper.

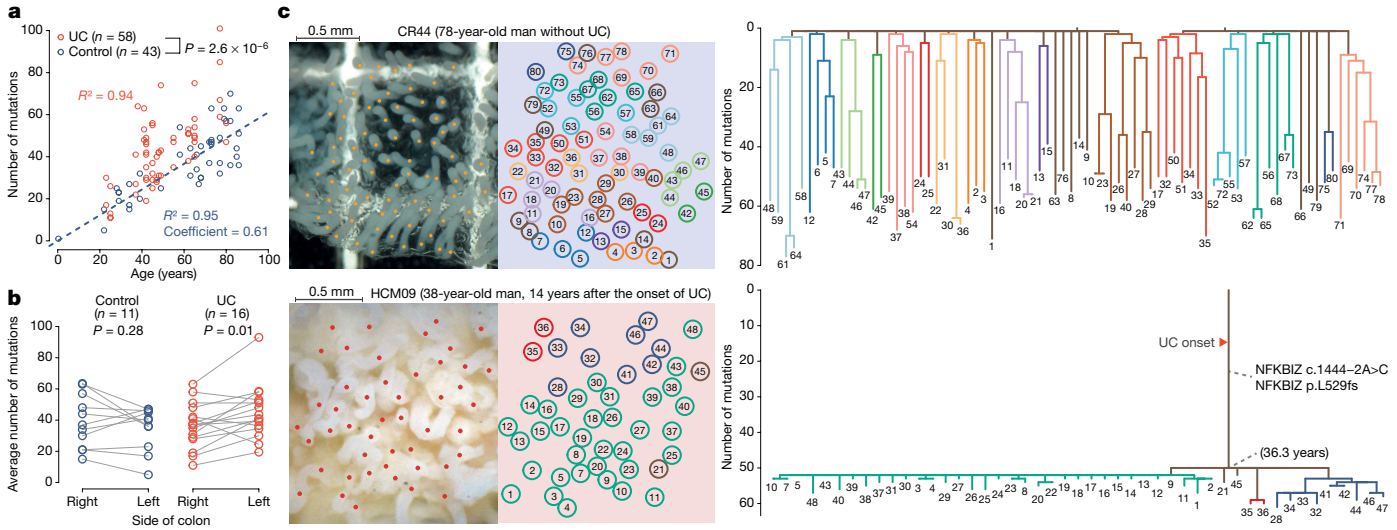


Fig. 1 | Somatic mutations in single crypts. a, Correlation between the number of mutations in single crypts (derived from WES) and the age of the individuals from whom the crypts were derived. Two-sided Mann–Whitney *U* test for comparison between UC and non-UC samples. A regression line that assumes an intercept of zero for non-UC samples is shown with R^2 and coefficient values (blue). The R^2 value (red) of the linear regression model for UC samples is also shown (Methods). **b**, Average number of mutations in paired single crypts from the right and left side of the colon in each patient with UC or control individual. Two-sided paired *t*-test. **c**, Clonal structure within adjacent crypts in epithelia from a control individual (CR44) and a patient with UC (HCM09). The corresponding clones in the phylogenetic trees and geographical maps are depicted by numbers and colours. The estimated time of UC onset, the age of the most-recent common ancestral crypt in HCM09 and branches that contain driver mutations are indicated (see Methods).

Clonal history of normal and UC crypts

The colorectal epithelium is composed of single-layer columnar cells that are compacted into numerous small replication units known as crypts. Each crypt comprises around 2,000 cells, which are replenished by the stem cells located at the bottom of the crypt[10]. During life, expansion of the colorectal epithelium and repair of physiological or pathological deficits of crypts are accomplished by crypt fission[11]. Thus, to understand the dynamics of UC and non-UC epithelia, we first evaluated somatic mutations in single crypts that were isolated from samples from patients with UC and unaffected individuals, using whole-exome sequencing (WES) of whole-genome-amplified single-crypt DNA (Extended Data Figs. 1–3). We found that the number of mutations in non-UC crypts increases with age at an annual mutation rate of 0.61 mutations per exome per year, and that this rate is significantly increased in crypts derived from patients with UC (Fig. 1a). When corrected for age at diagnosis and disease duration (Methods), the mutation rate in UC crypts was more than three times higher than that of non-UC crypts (1.9 mutations per exome per year). Whereas the mutation rate did not differ between the right and left side of the colon in control individuals, we observed that mutation rates were higher for the crypts from the left side of the colon in patients with UC, in which inflammation tends to be more severe than in the right side[8] (Fig. 1b). Although we found no significantly mutated genes among crypts from normal epithelia, three genes (*NFKBIZ*, *PIGR* and *ARID1A*) were significantly affected by non-synonymous mutations in UC epithelium (ratio of non-synonymous to synonymous substitutions (dN/dS) > 1.0; $q < 0.05$), suggesting the driver roles of these genes.

Next, we analysed the history of crypt expansion in normal and inflamed epithelia by isolating a cluster of crypts from a small area of colorectal epithelium (0.25–4 mm^2) and analysing them by WES, retaining their geographical information (Extended Data Fig. 2b). We analysed crypt clusters from three patients with UC (aged 38–56 years) and three unaffected individuals (aged 78–88 years). As seen in isolated crypts, UC crypt clusters had substantially higher mutation rates (2.86–15.5 mutations per exome per year) compared with non-UC crypt clusters (0.63–0.75 mutations per exome per year). Most of the crypts (99.4%) in both patients with UC and unaffected individuals shared one or more mutations with other crypts. On the basis of this finding, we constructed phylogenetic trees and superimposed them on the geographical maps of the crypts (Fig. 1c, Extended Data Fig. 4a–d). In non-UC epithelium, the majority (99/143, 69.2%) of the branch points in the phylogenetic trees were estimated to occur before 20 years of age (corresponding to 12.2 mutations) (Extended Data Fig. 4e). This suggests that the number of crypts steadily increases until the age of 20 years to expand the size of the intestine (with an annual fission rate of around 0.136), whereas in the steady state that occurs throughout the rest of life, crypt fissions occur only occasionally to balance the physiological deficits of crypts (0.00413 fissions per year) (Methods). By contrast, the crypts from patients with UC had very different phylogenetic structures (Fig. 1c, Extended Data Fig. 4c, d). The phylogenetic trees were characterized by a notably long trunk containing known driver mutations, which terminated in many short branches. This indicates that after repetitive destruction and regeneration, these UC crypts rapidly expanded from a positively selected common ancestor crypt very recently (during the past 1.7–4.7 years; 3.3 years on average).

Positive selection in the UC epithelium

To further investigate clonal expansion in the UC epithelium, we performed WES of bulk crypts isolated from 269 non-dysplastic samples that were collected from 22 patients with UC and 23 control individuals, with a median of 6 (range, 4–8) samples per individual (Extended Data Fig. 1b). We also analysed an additional 294 bulk-crypt samples, which were collected from surgically obtained rectums from 6 patients with UC at regular intervals (Extended Data Fig. 1b). Compared with non-UC samples, UC samples had much higher numbers of mutations (51.2 versus 1.9 mutations per sample on average) (Extended Data Fig. 5a). Maximum mutant-cell fractions (MCFs) were significantly higher for UC samples (mean = 0.58 (range, 0–1)) than non-UC samples (mean = 0.21 (range, 0–1)) ($P < 0.01$) (Methods), with 22.6% of UC samples showing a maximum MCF of 1.0. The number of mutations and the maximum MCF in UC samples weakly but significantly correlated with the interval between diagnosis and sampling (Extended Data Fig. 5b, c). Samples

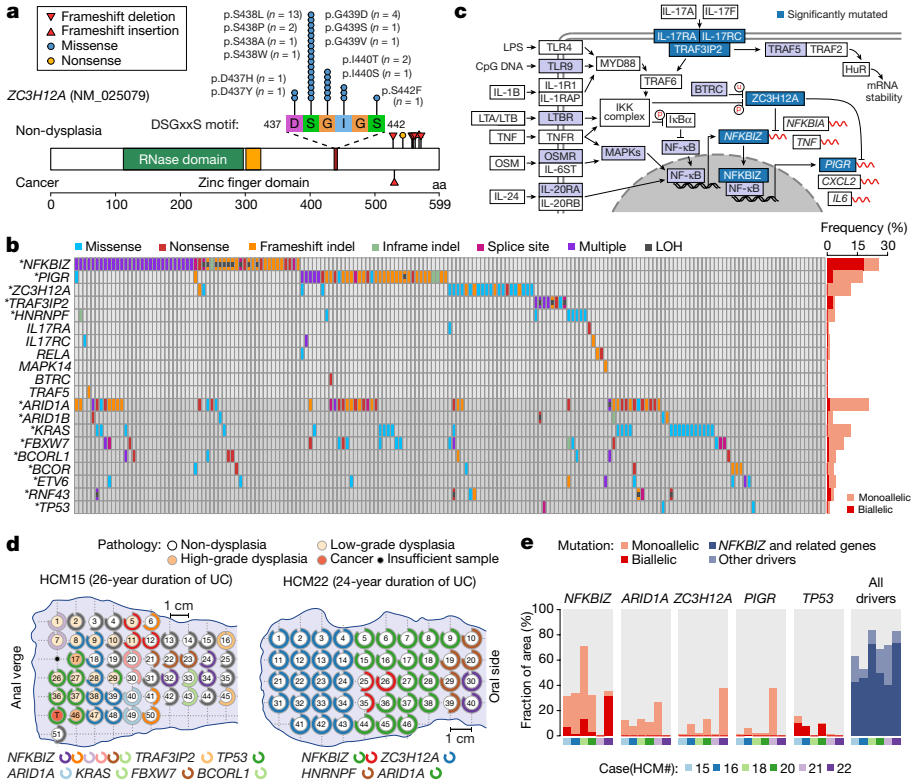


Fig. 2 | Driver genes and clonal expansion in the UC epithelium. **a**, Distribution of *ZC3H12A* mutations in non-dysplastic UC samples (58 isolated single crypts and 399 bulk crypts; top) and CAC samples ($n = 99$; bottom). **b**, Landscape of driver mutations that were detected in clones ($n = 183$) in non-dysplastic samples from patients with UC. The frequencies of each mutation are indicated on the right. Genes with asterisks are drivers in non-dysplastic samples from patients with UC. LOH, loss of heterozygosity. **c**, Mutations in IL-17 and other pro-inflammatory signalling pathways in non-dysplastic samples from patients with UC. Mutated components are indicated by coloured boxes. **d**, Expansion of clones within the rectum from two representative patients with UC (out of six). Histology is shown by the colour of the centre of each circle. Samples that contain independent clones are discriminated by the colour of the arch, and the size of the arch indicates the fraction of the largest clone (estimated by PyClone). Samples with a grey arch are an exception; none of the mutations in grey-arched samples are shared by other samples (including others with a grey arch). The driver mutations in the main trunks in each independent clone are indicated at the bottom, except in grey-arched samples, for which just the presence or absence of mutations in the IL-17-pathway is indicated (dark or light grey, respectively). **e**, Fraction of the total sampling area that is occupied by clones with the indicated mutations in each rectum derived from patients with UC.

from the left side of the colon tended to show higher numbers of mutations and higher MCFs compared with those from the right side (Extended Data Fig. 5d, e).

To decipher the mechanism of positive selection in the UC epithelium, we analysed the dN/dS ratios for 30,062 mutations that were detected in 399 non-dysplastic samples that were derived from patients with UC, and identified 14 driver genes for which there was evidence of positive selection (dN/dS > 1.0, $q < 0.05$) (Fig. 2a, b, Extended Data Fig. 6) (Methods). By contrast, no positively selected mutations were detected in 179 samples from control individuals. Most of these driver gene mutations in the UC epithelium were predicted to cause protein truncation and loss of function. Exceptions were mutations in *ZC3H12A*, *KRAS* and *HNRNPF*, which exhibited prominent mutational hotspots indicative of some neomorphic functions (Fig. 2a, Extended Data Fig. 6a). *NFKBIZ*, *TRAF3IP2* and *RNF43* frequently underwent biallelic inactivation, which was less common in other drivers (Fig. 2b).

A single bulk-crypt sample may contain multiple independent clones and subclones that carry discrete driver mutations. Thus, to precisely determine the frequency of mutations in the UC epithelium on an independent clone basis, we first clustered mutations that were detected in bulk-crypt samples using PyClone and—on the basis of the predicted cell fraction for each cluster of mutations—identified a total of 183 independent clones according to the 'pigeonhole principle' (Extended Data Fig. 1d, Methods). Among these, 166 (90%) contained one or more driver mutations, which most frequently affected *NFKBIZ*, followed by *ARID1A*, *PIGR*, *KRAS* and *ZC3H12A* (Fig. 2b). Notably, the products of the genes that were mutated in the UC epithelium overlapped with signalling pathways that converge on NFKBIZ (also known as IκBζ) and related molecules (ZC3H12A and PIGR) (Fig. 2c), with the mutations in these signalling-related molecules occurring in a mutually exclusive manner (Fig. 2b). Most prominent among these mutational targets are molecules that are involved in IL-17 signalling—including IL-17 receptors (IL17RA and IL17RC) and the adaptor of IL-17 receptors, TRAF3IP2, which were shown to be positively selected in the UC epithelium (Methods)—suggesting a major role of IL-17 signalling in this positive selection (Fig. 2c). Moreover, other inflammation-related cytokine receptors and pattern recognition receptors (PRRs) that are implicated in the upregulation of NFKBIZ (including OSMR, LTBR, IL20RA and TLR9[12]; Extended Data Fig. 6b) also underwent loss-of-function mutations in occasional clones (Methods).

We assessed the expansion of clones carrying mutations in these cytokine and PRR pathway components in surgically obtained rectums from six patients with UC using densely collected samples (Fig. 2d, Extended Data Fig. 7a). We identified 41 independent driver-mutated clones that involved two or more samples in these specimens, of which 31 had mutations in the cytokine and PRR pathways. Notably, in these patients, 53–83% of the entire rectum was made up of clones with one or more driver mutations, most of which were explained by mutations in the cytokine and PRR pathways (Fig. 2e). The clonal expansion could be extensive, occupying an epithelial area as large as 19 cm^2 (Extended Data

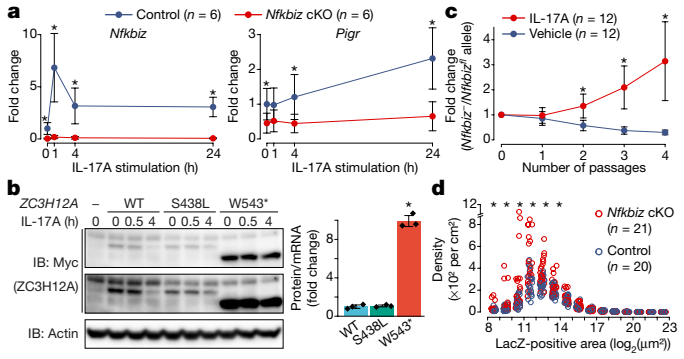


Fig. 3 | Effect of driver mutations. a, Fold change in expression levels of *Nfkbiz* and *Pigr* (mean ± s.d.) in colon epithelial organoids that were established from *Nfkbiz*$^{fl/fl}$ (control) and *Nfkbiz*$^{fl/fl}$*Vil1-cre* (*Nfkbiz* cKO) mice at the indicated time points after stimulation with IL-17A. Asterisks indicate $P < 0.05$ (two-sided Mann–Whitney *U* test). **b**, Left, immunoblot (IB) analysis of HCT116 cells that express empty vector, Myc-tagged wild-type (WT) *ZC3H12A* or the indicated *ZC3H12A* mutants after stimulation with IL-17A for the indicated times, using anti-Myc and β-actin antibodies. Data are representative of two independent experiments. Right, the ratio of protein to mRNA for ZC3H12A ($n = 3$ biological replicates) (mean ± s.d.). Asterisk indicates $P < 0.05$ between mutant and wild-type ZC3H12A (two-sided Student's *t*-test). For gel source data, see Supplementary Fig. 1. **c**, Fold change (mean ± s.d.) in the ratio of *Nfkbiz*$^{-}$ (knockout) to *Nfkbiz*^{fl} (flox) alleles at the indicated number of passages in *Nfkbiz* cKO and control organoids co-cultured in the presence or absence of IL-17A. Asterisks indicate $P < 0.05$ (two-sided Mann–Whitney *U* test). **d**, Density of LacZ-positive areas having indicated sizes in control and *Nfkbiz* cKO mice. Asterisks indicate $P < 0.05$ (two-sided Mann–Whitney *U* test).

Fig. 7b). In patient HCM22, almost all of the rectal epithelium (around 300,000 crypts) was replaced by only two clones that carry mutations in *NFKBIZ* (frameshift mutation at S339 (S339fs)) and *ZC3H12A* (S438L). In an analysis of phylogenetic structure, 37 clones had one or more driver mutations in their main trunk, and 26 of these 37 clones had mutations in the IL-17 pathway or other cytokine or PRR pathways—suggesting that these mutations were acquired before or early in the development of UC (Fig. 2d, Extended Data Fig. 7).

Effect of *NFKBIZ* and other mutations

Among these genes that are involved in cytokine and PRR signalling, those that were most commonly mutated were *NFKBIZ* (30%), *PIGR* (21%) and *ZC3H12A* (14%). NFKBIZ—which is upregulated by a number of primary stimuli, such as IL-17A, IL-1β and lipopolysaccharide (LPS)—is induced by activated NFκB to regulate the expression of secondary response genes that include those encoding many pro-inflammatory cytokines[13] (Fig. 2c, Extended Data Fig. 6b). *PIGR*, which encodes a key molecule for the transcytosis of immunoglobulin A (IgA)[14], is one such gene[15], and is known to be downregulated in the colorectal epithelium in IL-17R-deficient mice[15] and in patients with inflammatory bowel disease[16]. Thus, to understand the functional link between *NFKBIZ* and *PIGR* with regard to IL-17 signalling, we established *Nfkbiz*$^{-/-}$ and *Nfkbiz*$^{+/+}$ organoids from colorectal epithelium from *Nfkbiz*$^{fl/fl}$*Vil1-cre* and *Nfkbiz*$^{fl/fl}$ mice and examined the expression of *Pigr* after stimulation with IL-17A (Fig. 3a). In *Nfkbiz*$^{+/+}$ organoids, *Pigr* expression was induced by IL-17A slightly after the upregulation of *Nfkbiz* and increased until 24 h after stimulation. Conversely, in *Nfkbiz*$^{-/-}$ organoids, the induction of *Pigr* was significantly attenuated with no *Nfkbiz* induction (Fig. 3a). These observations confirm that NFKBIZ is a key regulator of *Pigr* expression upon IL-17A stimulation.

ZC3H12A encodes an RNase known as Regnase-1, which is implicated in the degradation of the mRNA of *NFKBIZ* and many downstream secondary response genes, including *PIGR*[17,18]. Mutations in the UC epithelium were highly enriched in the DSGxxS motif and the C-terminal part, and showed a prominent mutational peak at the S438 residue (Fig. 2a). This residue has been previously shown to be phosphorylated by the IκB kinase (IKK) complex to promote ubiquitination and subsequent degradation of the protein[19]. In mice, a mutation at the corresponding site, S435A, is reported to prevent ubiquitin-mediated degradation and lead to increased expression of ZC3H12A after stimulation with IL-1β[19]. We also demonstrated that cells that express *ZC3H12A*^{S438L} showed a higher level of protein expression compared with wild-type transfected cells (Fig. 3b, Extended Data Fig. 6c). Of note, we found a UC sample that contained a nonsense mutation affecting *BTRC*, which encodes the E3 ligase that is implicated in the ubiquitination of ZC3H12A. Moreover, we demonstrated that expression of a dominant-negative form of BTRC (BTRC(ΔF)) stabilized ZC3H12A when stimulated with IL-17A or IL-1β[19] (Extended Data Fig. 6d)—providing further support for the role of DSGxxS mutants in positive selection. An increase in protein expression compared to the wild type was also demonstrated for a C-terminal-truncating ZC3H12A mutant (W543*). Notably, the W543* mutant showed a much higher level of protein expression than the S438L mutant even though their levels of mRNA were similar (Fig. 3b), which suggests that the C-terminally deleted protein has a higher stability than the wild type regardless of IL-17A stimulation. Another gene that is implicated in the mRNA degradation of pro-inflammatory cytokine genes[20], *HNRNPF*, showed mutational hotspots that affect highly conserved tyrosine-containing residues (amino acids 201–210) (Extended Data Fig. 6a), although its link to the IL-17 signalling pathway is unknown.

We next examined the role of IL-17 signalling in the positive selection of NFKBIZ-deficient cells in the UC epithelium. When an equal number of colon organoids from *Nfkbiz*$^{-/-}$ and *Nfkbiz*$^{fl/fl}$ mice were mixed and co-cultured with or without IL-17A, we observed an increasing allelic ratio of *Nfkbiz*$^{-}$ relative to *Nfkbiz*^{fl} in the presence of IL-17A. By contrast, in the culture without IL-17A, the ratio was decreased, which suggests that in the presence of IL-17A, *Nfkbiz*-null cells have a selective advantage over cells in which the expression of *Nfkbiz* is intact (Fig. 3c). We also investigated the driver role of *NFKBIZ* in UC in vivo by generating *R26*$^{LacZ/WT}$*Nfkbiz*$^{fl/fl}$*Vil1-cre*^{ER} and *R26*$^{LacZ/WT}$*Nfkbiz*$^{WT/WT}$*Vil1-cre*^{ER} mice in which colitis was induced with dextran sulfate sodium (DSS); aberrant IL-17 signalling has previously been implicated in this DSS model of colitis[21]. In *R26*$^{LacZ/WT}$*Nfkbiz*$^{fl/fl}$*Vil1-cre*^{ER} mice, cells in which *Nfkbiz* was deleted were randomly introduced in the colorectal epithelium and visualized by LacZ staining. After treatment with tamoxifen (to induce Cre-mediated recombination), we induced chronic colitis in the mice by oral administration of DSS, and compared the size of the LacZ-positive area between *Nfkbiz*$^{-/-}$ and *Nfkbiz*$^{+/+}$ mice 11–12 weeks after tamoxifen was administered (Extended Data Fig. 8). Overall, the total fraction of the LacZ-positive areas, as well as the severity of inflammation (as evaluated by body weight, colon length and histology), did not significantly differ between these mice (Extended Data Fig. 8f–h). However, the number of LacZ-positive areas of size 256–16,384 μm^2 was significantly increased in *Nfkbiz*$^{-/-}$ mice compared with *Nfkbiz*$^{+/+}$ mice (Fig. 3d), which was not observed without treatment with DSS (Extended Data Fig. 8d). RNA sequencing of DSS-treated colorectal epithelium revealed downregulation of IL-17 signalling pathway genes and downstream targets of NFKBIZ in *Nfkbiz*$^{-/-}$ mice compared with *Nfkbiz*$^{+/+}$ mice (Extended Data Fig. 8i), which further supports the role of *NFKBIZ* mutations in the positive selection of cells in the UC epithelium.

Negative role of *NFKBIZ* mutations in CAC

We next investigated driver mutations and copy-number alterations (CNAs) in 99 CAC samples from patients with UC—including 65 samples for which WES data were publicly available[22–25] and 34 that were newly analysed with WES—together with 356 cases of sporadic colorectal cancer (sCRC) from The Cancer Genome Atlas (TCGA) (Fig. 4, Extended

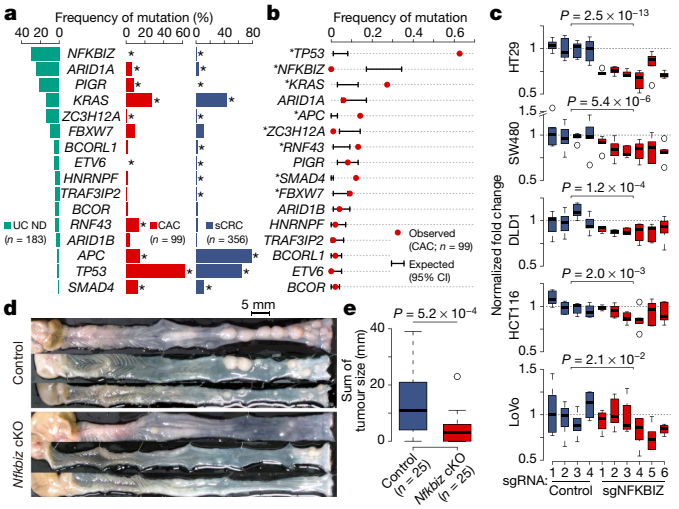


Fig. 4 | Comparison of driver mutations between UC non-dysplasia and cancer. a, Frequency of driver mutations among clones identified in non-dysplastic (ND) samples from patients with UC, as well as CAC and sCRC samples. Asterisks indicate $q < 0.1$ (two-sided Fisher's exact test with Benjamini–Hochberg adjustment). **b**, Observed frequency of mutation of the indicated genes in CAC samples from **a** and the 95% CI of the frequency of mutation expected from the observed fraction in 432 non-cancer samples from patients with UC. Asterisks indicate $q < 0.05$ (two-sided binomial test with Benjamini–Hochberg adjustment). **c**, Box plots of normalized fold change in the frequency of indicated sgRNA sequence reads after four weeks of culture compared with the pre-culture value in five human CRC cell lines ($n = 6$ biological replicates). Significant differences between control and *NFKBIZ*-directed sgRNAs are provided (two-sided Mann–Whitney *U* test). **d**, Representative images of AOM–DSS-induced colorectal tumours in control (3 out of 25; top) and *Nfkbiz* cKO (3 out of 25; bottom) mice. **e**, Box plots of sum of tumour size (that is, the sum of the maximum diameter of all tumours) per mouse. Two-sided Mann–Whitney *U* test. In all box plots, the median, first and third quartiles are indicated and whiskers extend to the furthest value within 1.5 × the interquartile range (IQR).

Data Figs. 1c, 9). We also performed WES in 13 samples with low- and high-grade dysplasia from patients with UC.

Dysplastic and CAC samples from patients with UC had significantly higher numbers of mutations compared with non-dysplastic samples from patients with UC (Extended Data Fig. 9c, d). An analysis of mutational signatures in UC samples identified two predominant signatures, signatures A and B (Methods) (Extended Data Fig. 9c, e). Corresponding to COSMIC signature 1, signature A mostly consisted of C>T substitutions in the CpG context and was predominant across in all tissue types. Signature B, which corresponds to COSMIC signature 17, was characterized by T>G and T>C substitutions in a C[T]T context and was highly enriched in UC samples (Extended Data Fig. 9f), implying that it is linked to chronic inflammation.

The profile of driver mutations was substantially different between non-dysplastic UC samples and cancer samples. dN/dS analysis identified six genes that were significantly positively selected in CAC (Extended Data Fig. 9a). These genes were also commonly mutated in sCRC (Fig. 4a) but were rarely affected in non-dysplastic UC samples, in which mutations in *NFKBIZ*, *ARID1A*, *ZC3H12A*, *PIGR* and *ETV6* were significantly enriched. Copy-number profiles were also very different between CAC or dysplastic samples, and non-dysplastic samples. CNAs were obligatory and extensive in CAC, but much less common (41%) in non-dysplastic samples (Extended Data Fig. 9a, g, h).

Although highly prevalent in non-dysplastic UC samples, *NFKBIZ* mutations were found in none of the 99 CAC samples. Thus, CAC do not develop from *NFKBIZ*-mutated cells, which suggests that *NFKBIZ*-mutated clones that were positively selected in UC epithelium have a strong relative disadvantage in CAC ($P \approx 7.7 \times 10^{-13}$) (Methods) (Fig. 4b). Similarly, *NFKBIZ* mutations were rarely seen in cases of sCRC from the TCGA (2/356) and in a cohort of Japanese patients with sCRC (6/525); among all cases from both of these cohorts, only one had a truncating mutation (Fig. 4a, Extended Data Fig. 9b)—also supporting the negative selection of *NFKBIZ*-mutated cells. To test this hypothesis, we transduced a lentiviral library consisting of six *NFKBIZ*-targeting single-guide (sg)RNA constructs and four non-specific (control) sgRNA constructs for CRISPR–Cas9-mediated insertions or deletions into five human colorectal cancer (CRC) cell lines, and assessed the enrichment of cells that were targeted by each sgRNA construct during four weeks of cell culture (Extended Data Fig. 10a). We observed a significant reduction in cell populations that were transduced with *NFKBIZ*-directed sgRNAs compared with cells that were transduced with control sgRNA in all CRC cell lines (Fig. 4c). Moreover, the number and size of colorectal tumours that were induced by inflammation using DSS and azoxymethane (AOM) were significantly decreased in *Nfkbiz*$^{fl/fl}$*Vil1-cre* mice compared with control *Nfkbiz*$^{fl/fl}$ mice (Fig. 4d, e, Extended Data Fig. 10b–f). These results of functional analyses again corroborate the negative selection of *NFKBIZ*-mutated CRC cells. A significant reduction in the frequency of *ZC3H12A* mutations in CAC compared to UC epithelium was also observed, in sharp contrast to *TP53* and *KRAS* mutations, which were substantially enriched in CAC (Fig. 4b).

Discussion

In the steady state of the intestine that occurs in adult life, crypts in the colon still divide to balance any physiological deficits of crypts. Assuming a constant number of crypts in the adult colon, 50,000 crypts are estimated to be lost annually and replaced by new ones through crypt fissions—equivalent to around 0.4% of the total colorectal crypts per year. In patients with UC, the rate of crypt fission is markedly accelerated (around 2.16 fissions per crypt per year) to correct the larger deficits of crypts; strong positive selection operates to bring about the dominance of clones that carry mutations in the IL-17 and other cytokine and PRR pathways. The importance of these pathways in the pathogenesis of UC is supported by GWAS studies that report a significant association of UC with the *NFKBIZ* and *TRAF3IP2* loci[26,27].

Despite a substantial overlap of affected driver genes between CAC and UC epithelia, their frequencies significantly differed between the two tissues, indicating that discrete mechanisms of positive selection exist in cancerous and non-cancerous tissues. Our results from the UC epithelium suggest that the mechanism that allows for positive selection under chronic inflammation may not necessarily contribute to neoplastic growth. On the contrary, it may impose strong negative effects on the development of cancer, as is evident from our finding that *NFKBIZ* mutations are extremely rare in CAC despite being pervasive in non-dysplastic samples of UC.

Finally, one of the central implications of our study is the possibility that chronic inflammation can evoke a tissue response by eliciting the remodelling of tissues through positive selection of clones that acquire mutations—to the extent that the entire rectum becomes almost completely reconstituted by a few clones that carry IL-17-pathway-related mutations. Chronic inflammation is highly prevalent among the human population and such a mechanism of tissue remodelling through positive selection may potentially modify the picture of disease in affected organs and tissues depending on the nature of inflammation. A comprehensive understanding of such responses will be important for improving our knowledge of the pathogenesis of inflammatory diseases and for the development of novel therapeutics.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information,

acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1856-1.

1. Grivennikov, S. I., Greten, F. R. & Karin, M. Immunity, inflammation, and cancer. *Cell* **140**, 883–899 (2010).
2. Hunter, P. The inflammation theory of disease. The growing realization that chronic inflammation is crucial in many diseases opens new avenues for treatment. *EMBO Rep.* **13**, 968–970 (2012).
3. Nathan, C. & Ding, A. Nonresolving inflammation. *Cell* **140**, 871–882 (2010).
4. Jaiswal, S. et al. Age-related clonal hematopoiesis associated with adverse outcomes. *N. Engl. J. Med.* **371**, 2488–2498 (2014).
5. Martincorena, I. et al. Tumor evolution. High burden and pervasive positive selection of somatic mutations in normal human skin. *Science* **348**, 880–886 (2015).
6. Yokoyama, A. et al. Age-related remodelling of oesophageal epithelia by mutated cancer drivers. *Nature* **565**, 312–317 (2019).
7. Burisch, J., Jess, T., Martinato, M. & Lakatos, P. L. The burden of inflammatory bowel disease in Europe. *J. Crohns Colitis* **7**, 322–337 (2013).
8. Ungaro, R., Mehandru, S., Allen, P. B., Peyrin-Biroulet, L. & Colombel, J. F. Ulcerative colitis. *Lancet* **389**, 1756–1770 (2017).
9. Eaden, J. A., Abrams, K. R. & Mayberry, J. F. The risk of colorectal cancer in ulcerative colitis: a meta-analysis. *Gut* **48**, 526–535 (2001).
10. Kang, H. & Shibata, D. Direct measurements of human colon crypt stem cell niche genetic fidelity: the role of chance in non-Darwinian mutation selection. *Front. Oncol.* **3**, 264 (2013).
11. Choi, C. R., Bakir, I. A., Hart, A. L. & Graham, T. A. Clonal evolution of colorectal cancer in IBD. *Nat. Rev. Gastroenterol. Hepatol.* **14**, 218–229 (2017).
12. Eto, A., Muta, T., Yamazaki, S. & Takeshige, K. Essential roles for NF-κB and a Toll/IL-1 receptor domain-specific signal(s) in the induction of IκB-ζ. *Biochem. Biophys. Res. Commun.* **301**, 495–501 (2003).
13. Yamamoto, M. et al. Regulation of Toll/IL-1-receptor-mediated gene expression by the inducible nuclear protein IκBζ. *Nature* **430**, 218–222 (2004).
14. Bruno, M. E. C., Frantz, A. L., Rogier, E. W., Johansen, F. E. & Kaetzel, C. S. Regulation of the polymeric immunoglobulin receptor by the classical and alternative NF-κB pathways in intestinal epithelial cells. *Mucosal Immunol.* **4**, 468–478 (2011).
15. Cao, A. T., Yao, S., Gong, B., Elson, C. O. & Cong, Y. Th17 cells upregulate polymeric Ig receptor and intestinal IgA and contribute to intestinal homeostasis. *J. Immunol.* **189**, 4666–4673 (2012).
16. Bruno, M. E. et al. Correlation of biomarker expression in colonic mucosa with disease phenotype in Crohn's disease and ulcerative colitis. *Dig. Dis. Sci.* **60**, 2976–2984 (2015).
17. Nakatsuka, Y. et al. Pulmonary Regnase-1 orchestrates the interplay of epithelium and adaptive immune systems to protect against pneumonia. *Mucosal Immunol.* **11**, 1203–1218 (2018).
18. Mino, T. et al. Regnase-1 and Roquin regulate a common element in inflammatory mRNAs by spatiotemporally distinct mechanisms. *Cell* **161**, 1058–1073 (2015).
19. Iwasaki, H. et al. The IκB kinase complex regulates the stability of cytokine-encoding mRNA induced by TLR-IL-1R by controlling degradation of regnase-1. *Nat. Immunol.* **12**, 1167–1175 (2011).
20. Reznik, B., Clement, S. L. & Lykke-Andersen, J. hnRNP F complexes with tristetraprolin and stimulates ARE-mRNA decay. *PLoS One* **9**, e100992 (2014).
21. Ramirez-Carrozzi, V. et al. IL-17C regulates the innate immune function of epithelial cells in an autocrine manner. *Nat. Immunol.* **12**, 1159–1166 (2011).
22. Robles, A. I. et al. Whole-exome sequencing analyses of inflammatory bowel disease-associated colorectal cancers. *Gastroenterology* **150**, 931–943 (2016).
23. Fujita, M. et al. Genomic landscape of colitis-associated cancer indicates the impact of chronic inflammation and its stratification by mutations in the Wnt signaling. *Oncotarget* **9**, 969–981 (2018).
24. Din, S. et al. Mutational analysis identifies therapeutic biomarkers in inflammatory bowel disease-associated colorectal cancers. *Clin. Cancer Res.* https://doi.org/10.1158/1078-0432.CCR-17-3713 (2018).
25. Baker, A. M. et al. Evolutionary history of human colitis-associated colorectal cancer. *Gut* **68**, 985–995 (2019).
26. Jostins, L. et al. Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. *Nature* **491**, 119–124 (2012).
27. Liu, J. Z. et al. Association analyses identify 38 susceptibility loci for inflammatory bowel disease and highlight shared genetic risk across populations. *Nat. Genet.* **47**, 979–986 (2015).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

[1]Department of Pathology and Tumour Biology, Kyoto University, Kyoto, Japan. [2]Institute for the Advanced Study of Human Biology (WPI-ASHBi), Kyoto University, Kyoto, Japan. [3]Department of Gastroenterology and Hepatology, Kyoto University, Kyoto, Japan. [4]Department of Inflammatory Bowel Disease, Division of Surgery, Hyogo College of Medicine, Nishinomiya, Japan. [5]Department of Surgical Pathology, Hyogo College of Medicine, Nishinomiya, Japan. [6]Department of Medical Chemistry, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [7]Department of Surgery, Kyoto University, Kyoto, Japan. [8]Gastroenterological Center, Department of Gastroenterological Surgery, Cancer Institute Hospital, Japanese Foundation for Cancer Research, Tokyo, Japan. [9]Department of Therapeutic Oncology, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [10]Department of Gastroenterology and Hepatology, Kyorin University School of Medicine, Tokyo, Japan. [11]Department of Diagnostic Pathology, Kyoto University Hospital, Kyoto, Japan. [12]Department of Diagnostic Pathology, Kyoto University, Kyoto, Japan. [13]Division of Molecular Oncology, National Cancer Centre Research Institute, Tokyo, Japan. [14]Laboratory of DNA Information Analysis, Human Genome Centre, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [15]Laboratory of Sequence Analysis, Human Genome Centre, Institute of Medical Science, The University of Tokyo, Tokyo, Japan. [16]Department of Advanced Diagnosis, Clinical Research Centre, National Hospital Organization Nagoya Medical Center, Nagoya, Japan. [17]Research and Development Centre for Precision Medicine, University of Tsukuba, Tsukuba, Ibaraki, Japan. [18]Department of Immunology, Akita University Graduate School of Medicine, Akita, Japan. [19]Division of Experimental Therapeutics, Kyoto University, Kyoto, Japan. [20]DSP Cancer Institute, Sumitomo Dainippon Pharma, Osaka, Japan. [21]Division of Hepato-Biliary-Pancreatic Surgery and Transplantation, Department of Surgery, Graduate School of Medicine, Kyoto University, Kyoto, Japan. [22]Department of Paediatric Surgery, Kanazawa Medical University, Ishikawa, Japan. [23]Department of Gastroenterology and Hepatology, Sapporo Medical University School of Medicine, Sapporo, Japan. [24]Kansai Electric Power Hospital, Osaka, Japan. [25]Department of Medicine, Centre for Haematology and Regenerative Medicine, Karolinska Institute, Stockholm, Sweden. *e-mail: sogawa-tky@umin.ac.jp

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment. Investigators were blinded to the genotypes of mice until the statistical evaluation.

Subjects and materials

We enrolled a total of 647 patients who underwent surgery or endoscopy of the lower gastrointestinal tract at Kyoto University Hospital, Hyogo College of Medicine Hospital and Cancer Institute Hospital. Among these, 76 had been diagnosed with UC and the remaining 571 were either diagnosed with other diseases such as colorectal polyp or cancer without associated colitis or turned out to be healthy individuals. This study was approved by the ethical committees at Kyoto University, Hyogo College of Medicine and Cancer Institute Hospital. We have complied with all relevant ethical regulations. Informed consent was obtained from all participants. Characteristics of the participants are summarized in Supplementary Table 1.

Specimens from individuals without UC were endoscopically or surgically obtained from non-dysplastic intestinal mucosa. After methylene blue staining, they were carefully assessed under a stereomicroscope to exclude samples involving dysplastic crypt lesions such as aberrant crypt foci. All specimens obtained from patients with UC were pathologically evaluated using half of each specimen or its adjacent mucosa. Among them, six surgically resected rectum samples were comprehensively examined by 5-mm sections across the whole sampling area in 1-cm lattices. On the basis of Riddell's classification[28], pathological evaluation was performed by two expert pathologists (Y.T. and T. Sakurai, or T.K. and S.H.) and samples were classified into four groups: non-dysplasia, low-grade dysplasia, high-grade dysplasia and cancer.

To obtain single crypts and crypts in bulk, samples obtained by endoscopic biopsy were trimmed to a size of approximately 4 mm^2. From surgical specimens, samples were collected by punch biopsy of 2.5 mm in diameter (4.9 mm^2) (BP-25F, Kai Industries). After being separated from the submucosal layers by incubating samples in 20 mM ice-cold EDTA in phosphate-buffered saline (PBS) for 20 min, the epithelium was mechanically dissociated using fine forceps, and single-crypt or bulk-crypt samples were collected (Extended Data Fig. 2a). To obtain a geographically mapped cluster of crypts, two methods were used. For non-UC samples, the crypt structure within the mucosa was first fixed with acrylic adhesive from the epithelial surface, then the underneath cluster of crypts was peeled off en bloc from the submucosal layers after treatment with 20 mM EDTA. For UC samples, a gentle pressure was applied to the whole mucosal tissue from both sides to push up crypts after treatment with 20 mM EDTA. Finally, individual crypts were isolated and their position was recorded and numbered (Extended Data Fig. 2b). For the analysis of cancer samples, tumour cells were freshly obtained or dissected from formalin-fixed paraffin-embedded (FFPE) tissue under a stereomicroscope and the DNA was extracted.

For WES, genomic DNA isolated from each single crypt was split into two aliquots, each of which was independently subjected to whole-genome amplification (WGA) with the REPLI-g Single Cell Kit (Qiagen) and used for either WES or subsequent validation. For other freshly obtained samples, genomic DNA and RNA were extracted simultaneously using the AllPrep DNA/RNA Micro Kit (Qiagen). DNA was extracted from FFPE cancer samples using the GeneRead DNA FFPE Kit (Qiagen). DNA from peripheral blood or tumour-free muscularis propria that was macro-dissected from FFPE cancer samples was extracted using the Gentra Puregene Kit (Qiagen) or GeneRead DNA FFPE Kit (Qiagen), respectively, and used as a germline control. For targeted capture sequencing of sCRCs, genomic DNA was extracted from cryo-preserved samples using the AllPrep DNA/RNA Mini Kit (Qiagen). Sample information is summarized in Supplementary Tables 2 and 3.

Whole-exome sequencing

WES libraries were prepared using SureSelect Human All Exon V5 (Agilent Technologies) or xGen Exome Research Panel (IDT), followed by sequencing of enriched exon fragments on a HiSeq 2500 or NovaSeq 6000 system (Illumina) in 100–150-bp paired-end mode as previously described[29]. The target depth was 100×, and the actual depth was 133× (60–293×). The mean depth in germline controls from 122 individuals was 140×(55–204×). Mutation calling was performed using the Genomon2 pipeline (v.2.6), as previously described[6]. In brief, sequencing reads were aligned to the human reference genome (GRCh37) using Burrows-Wheeler Aligner (v.0.7.8) with default parameter settings. PCR duplicates were eliminated using biobambam (v.0.0.191). Somatic mutations were detected by eliminating polymorphisms and sequencing errors. To achieve this, Genomon2 first discards any low-quality, unreliable reads and variants according to the following criteria: (i) mapping quality < 20, (ii) base call quality < 15. After further excluding those variants that are not supported by (i) a sufficient number of reads (total reads ≥ 8 and variant reads ≥ 3); (ii) variant allele frequencies (VAFs) ≥ 0.05 (for fresh and FFPE samples), ≥ 0.25 (for single-crypt WGA samples) and < 0.02 (for germline control); (iii) a strand ratio not equal to 0 or 1, the remaining variants were interrogated to establish whether they were observed at significantly higher VAFs than expected for errors ($P \leq 10^{-4}$, $P \leq 10^{-3.5}$ and $P \leq 10^{-4}$ for fresh, FFPE and WGA samples, respectively). Significance was evaluated by the EBCall algorithm[30], on the basis of an empirical distribution of VAFs as determined using WES data of non-paired peripheral blood samples ($n = 20$). Putative germline variants were also excluded by comparing VAFs with matched controls using the Fisher's exact test ($P \leq 10^{-1}$, $P \leq 10^{-1}$ and $P \leq 10^{-2}$ for fresh, FFPE and WGA samples, respectively), which also eliminated any remaining sequencing errors. The monoclonal origin of single-crypt cells, which are derived from a single stem cell at a given time, was evident from the normal distribution of VAFs around 0.5 (Extended Data Fig. 3a, b).

For mapping samples of clusters of crypts and surgically resected rectums, each of the mutations detected in a sample was interrogated in all of the remaining samples. If a variant detected in one or more samples by the above mutation call was also present (VAF ≥ 0.1 and ≥ 0.05 for single-crypt mapping and rectum mapping, respectively) in another sample within the same mapping specimen, the variant was considered to be true positive, because it has a high prior probability for true positivity.

Validation of detected mutations

Validation of mutations detected by WES in fresh and FFPE samples was performed using PCR-based deep (>500×) sequencing as previously described[29], in which 708 single-nucleotide variants (SNVs) and 28 indels were randomly selected and evaluated. A mutation was considered to be validated when (i) the sequencing depth was ≥ 500× in both test and germline control samples; (ii) the VAF in the test sample was 5 times higher than that in the corresponding germline control sample; and (iii) the VAF in the test sample was high enough (≥ 0.01) to be captured by the lower threshold in WES (0.05)[6]. The overall validation rate was 99.0% for fresh samples (390/394 mutations) and 97.4% for FFPE samples (333/342 mutations). Validation of mutations detected in WGA DNA of single crypts was performed using WES of replicated samples (Extended Data Fig. 2c), in which 126 SNVs and indels detected in 3 single crypts were evaluated. A mutation was considered to be validated when (i) the sequencing depth in the replicated sample was ≥ 8; (ii) the VAF in the replicated sample was > 0.2. The validation rate of WGA DNA was 98.4% (124/126 mutations). The results of validation sequencing are summarized in Supplementary Table 4.

External dataset

WES data (bam files) of paired tumour and germline control samples from patients with sCRC were downloaded from the TCGA data portal. Bam files from the TCGA were converted to fastq format using biobambam and processed with the same pipeline for mutation calling as was applied to the fresh samples that were obtained in this study (see above). For accuracy, sequencing data generated from tumours using WGA were excluded. Hypermutated tumours (≥ 12 mutations per 10^6 bases) were also excluded from the comparison of driver genes between non-dysplastic samples from patients with UC and TCGA sCRC samples. Lists of mutations in 65 cases of colitis-associated cancer (CAC) in individuals with UC were available from four previous publications of UC–CAC WES studies[22–25]. Samples included in this study are summarized in Supplementary Table 2.

Analysis of copy-number alterations

CNAs were evaluated from WES data using our in-house pipeline, 'CNACS'[6]. The fractions of genomic regions with CNAs were compared between non-dysplasia, dysplasia and CAC in patients with UC. For expanded clones that were observed in more than one sample, the average fraction with CNAs was used.

Estimation of mutant-cell fractions

Precise estimation of the fraction of cells that have a somatic mutation (mutant-cell fraction (MCF)) might be complicated when the mutation involves a gene located on the chromosomal segment that shows a CNA. Most mutations detected by WES in isolated non-dysplastic samples were outside of regions with CNAs. Thus, for those mutations (7,024/7,072), the MCF was simply estimated from their VAFs as follows:

MCF = 2 × VAF (for autosomes and the X chromosome in females); or
MCF = VAF (for the X chromosome in males).

By contrast, when the mutated locus does show an abnormal copy number or allelic imbalance, estimation of the MCF depends on several factors that may not be inferred in a deterministic way—particularly in cases in which a sample contains multiple clones with different compositions of mutant alleles. However, for a mutation in a region that exhibits copy-number loss or allelic imbalance, using the following calculation does not lead to overestimation of the MCF. Thus, we estimate the MCF of such mutations (31/7,072) as follows:

MCF = 2 × VAF − %UPD/100 (for uniparental disomies; UPDs); or
MCF = VAF × (2 − %Del/100) (for deletions),

where %UPD and %Del denote the estimated fraction (%) of the respective abnormality. Other mutations associated with copy-number gain (17/7,072) were excluded from the MCF analysis in Extended Data Fig. 5.

Analysis of significantly mutated (driver) genes

Significantly mutated genes, or driver genes, were investigated on the basis of dN/dS, which was calculated for all genes using the dNdScv package in R. For surgically resected rectum specimens, for which samples were obtained according to 1-cm lattices, mutations from the same clone could be detected in multiple samples. Such observations of multiple mutations belonging to the same clone needed be corrected before the analysis. To achieve this, we first clustered samples into sets in which samples shared four or more mutations with VAF ≥ 0.25 (Supplementary Table 5), or otherwise treated them as single samples. Each mutation detected in a given case was included in either a sample set or a single sample that had the highest VAF of the mutation. After applying dNdScv to the mutations in non-dysplasia (36 sample sets including 152 samples, and 243 single samples) and non-hypermutated CAC (56 samples including our samples and publicly available data for which information on synonymous mutations was present) in patients with UC separately, those genes that were significantly mutated (dN/dS > 1, $q < 0.05$) and found in ≥ 4% of non-dysplasia or CAC samples were considered as drivers (14 genes in non-dysplasia; 6 genes in CAC) (Supplementary Table 6). Driver genes were also analysed using MutSigCV, which confirmed 10/14 and 4/6 genes as being significantly mutated ($q < 0.05$) in non-dysplasia and cancer in patients with UC, respectively.

In UC non-dysplasia, we investigated loss-of-function mutations in genes annotated as cytokine receptor ($n = 108$) or PRR ($n = 33$) genes. Through WES of 399 non-dysplastic samples from 28 patients with UC, we identified a total of 2,575 truncating mutations (including those affecting splice sites). The significance of the mutations in each gene was evaluated by calculating the type-I error rate, assuming a Poisson distribution with observed mutation rate for truncating mutations (2,575/33,479,141 bp) and the length of each gene. Multiple testing was corrected using the Benjamini–Hochberg method. Genes with $q < 0.05$ were considered to be significantly mutated. We also calculated dN/dS values for these 141 genes. Information on the cytokine receptor and PRR genes is summarized in Supplementary Table 7. Genes with a protein-truncating mutation that affects a component of the IL-17 signalling pathway were considered to be potential drivers. In total, 20 genes were considered to be related to positive selection in UC non-dysplasia (Fig. 2b).

Targeted capture deep sequencing of driver genes

Targeted capture deep sequencing of 525 non-hypermutated sCRCs from a Japanese cohort was performed according to the manufacturer's protocol, using a custom bait library (xGen Predesigned Gene Capture Pools; IDT) that was designed to capture 16 genes that include drivers in either UC non-dysplasia or CAC (Supplementary Table 6). After hybridization capture, the enriched DNA fragments were sequenced on a HiSeq 2500 system. The mean depth was 1,879× (364–2,545×). Relevant somatic mutations were called using the Genomon2 pipeline and were filtered by excluding (i) synonymous SNVs; (ii) EBCall P value > 10^{-4}; (iii) VAF < 0.02; (iv) strand ratio = 0 or 1; (v) known variants listed in SNP databases (1000 Genome Project, Exome Sequencing Project 6500, Human Genetic Variation Database and Exome Aggregation Consortium). Candidate mutations were further filtered by removing missense SNVs with a VAF of 0.4–0.6, except for SNVs that have 5 or more entries in the Catalogue of Somatic Mutations in Cancer (COSMIC) (v.70). Detected mutations are summarized in Supplementary Table 8.

Mutational signature

Mutational signatures in 29,409 SNVs—consisting of 1,537 non-UC-derived SNVs detected by WES of single crypts ($n = 43$); 16,428 SNVs of non-dysplastic samples from patients with UC, which were either directly detected from WES of single crypts ($n = 58$) or assigned to PyClone-imputed clones from bulk-crypt samples ($n = 183$); 2,022 SNVs assigned to clones from dysplastic samples from patients with UC ($n = 13$); and 9,422 SNVs detected in non-hypermutated CAC ($n = 92$)—were extracted using the R package pmsignature. Extracted signatures were compared with the COSMIC signatures using the R package deconstructSigs. Signature A was confirmed to be highly similar to COSMIC signature 1 and signature B to COSMIC signature 17. The frequencies of signature B were compared between non-dysplastic samples from control individuals, and non-dysplastic, dysplastic and CAC samples or clones from patients with UC. Samples or clones with fewer than 10 mutations were excluded from the analysis.

Detection of crypt expansion in the colorectal epithelium in patients with UC

To identify independent clones for isolated bulk-crypt samples ($n = 133$) from patients with UC, clonal composition in these samples was interrogated using PyClone (v.0.13.0), in which a set of clones that have a discrete set of mutations were imputed together with their estimated cellular prevalence. For samples for geographical mapping in the

rectum (n = 294; 49 samples per individual on average (range, 41–57) from six patients with UC), we first identified those samples that shared four or more mutations with VAF ≥ 0.25, which were considered to contain a clone expanded from the same ancestral crypt and therefore analysed as a group using PyClone. For other geographical samples, samples were analysed as isolated samples (Supplementary Table 5). When the maximum cellular prevalence of mutational clusters is more than 0.5, the Pigeonhole principle can be used to decide whether the dominant cell fraction includes other minor fractions of mutated cells or not. Thus, only samples with a maximum cellular prevalence ≥ 0.5 were included for the analyses of expanded crypts, and minor mutational clusters were included only when the sum of such dominant cell fractions and minor cell fractions exceeded 1.0 (Extended Data Fig. 1d). Driver mutations, and their clone size, that were detected in individual samples in the geographical mapping of six patients with UC are summarized in Supplementary Table 9.

Phylogenetic analysis

To analyse the evolutionary history of clusters of crypts by microscale mapping, genotype information at every locus that was mutated in one or more samples was used to construct a phylogenetic tree. The maximum parsimony method using the Subtree-Pruning-Regrafting algorithm was performed by MEGA X, and the most-parsimonious tree from 1,000 replicates was obtained. Time points for the onset of UC in phylogenetic trees were estimated from the age of UC onset and the rate of mutation rate in control individuals (Fig. 1a). The age of the most-recent common ancestral crypt in UC trees was estimated from the time point of UC onset, the height of the first branch point and the average number of mutations in analysed crypts within each patient with UC (assuming a linear accumulation of mutations after the onset of UC).

In the multi-sampling analysis for rectum samples, phylogenetic trees were constructed for those clones observed in two or more samples. The assignment of different subclones to distinct clonal structures was accomplished according to the Pigeonhole principle, on the basis of the cellular prevalence of each mutational cluster as estimated from PyClone (as described previously[6]). In brief, when and only when the sum of their cellular prevalence exceeded 1.0, the two clusters were thought to be in the same clonal structure. For clusters in which one had a smaller cellular prevalence, the cluster was assigned as a subclone of the other cluster with larger cellular prevalence. All clusters for which clonal structures were undetermined were excluded from further analyses. Representative phylogenetic trees are shown in Extended Data Fig. 7c–h.

Estimation of the rate of crypt fission

In control individuals, the rate of crypt fission (fissions per crypt per year) before 20 years of age was estimated on the basis of the number of ancestral crypts at birth and the number of crypts at 20 years of age in phylogenetic trees in three clusters of crypts, according to the following formula:

$$\text{Fission rate} = (\log_2(N_{\text{crypt}} \text{ at 20 years old}) - \log_2(N_{\text{crypt}} \text{ at birth}))/20,$$

where N_{crypt} is the number of crypts and where we assumed that: (i) every crypt had < 1 somatic mutation at the time of birth; (ii) the intestine continues developing until 20 years of age, during which time crypts will not be lost.

The fission rate after 20 years of age was estimated on the basis of the number of branches that occur after 20 years of age to the sampling time point, according to the following formula:

$$\text{Fission rate} = (N_{\text{branch point}}/N_{\text{crypt}})/Y_{\text{after 20}},$$

where $N_{\text{branch point}}$ is the number of branch points after 20 years of age (12.2 mutations) and $Y_{\text{after 20}}$ denotes years, with 20 subtracted from the age of the subject. We assumed that: (i) colon size does not change after 20 years of age; (ii) crypt fissions occur only to balance the physiological deficits.

In patients with UC, the fission rate was estimated on the basis of the rate of mutation after disease onset and the time point of the most-recent common ancestral branch point of analysed crypts, according to the following formula:

$$\text{Fission rate} = \log_2(N_{\text{crypt}})/Y_{\text{afterMRCA}},$$

where $Y_{\text{afterMRCA}}$ denotes the number of years after the most-recent common ancestral (MRCA) branch point to the age of the subject, and where we assumed that: (i) the rate of mutation after the disease onset was constant; (ii) crypts were not lost after the time point of the most-recent common ancestral crypt.

Comparison between the mutation frequency in CACs and the mutated fraction in UC epithelium

There was a substantial discrepancy in mutation frequencies in driver genes between CAC and non-dysplastic samples from patients with UC. Thus, to evaluate the positive or negative role of mutations in each driver gene in colitis-associated carcinogenesis, we compared the mutation frequency of each gene in 99 samples of CAC (Fig. 4a) with that expected from the fraction of the mutated area in UC colorectal epithelium under the null hypothesis H_0: the gene mutation has no selective effect on the development of CAC from background UC colorectal epithelium. First, for each driver gene, the mean fraction of mutated area ($F_{\text{UC_epithelia}}$) in non-cancer samples from patients with UC (n = 432) was calculated by using the model of cellular prevalence in PyClone. Then, the 95% confidence interval (95% CI) of the expected mutation frequency in CACs on the basis of the mutated fraction in UC colorectal epithelium was calculated for each gene, from a binomial distribution, $B(t, 99)$ using $F_{\text{UC_epithelia}}$ as an estimated value of prevalence (t) of mutated area. For each driver gene, the significance of the difference between the observed and the expected mutation frequency in CAC was calculated by two-sided binomial test. Multiple testing was corrected using the Benjamini–Hochberg method. Detailed results are summarized in Supplementary Table 10.

Cell culture

HeLa cells were maintained in DMEM (08459-64, Nacalai Tesque) supplemented with 10% FBS and 100 μM 2-mercaptoethanol (21418-42, Nacalai Tesque). HCT116, HT29, SW480, DLD1 and LoVo cells were maintained in DMEM supplemented with 10% FBS.

Introduction of wild-type and mutant *ZC3H12A*

cDNAs of full-length human wild-type *ZC3H12A* (encoding the RNase ZC3H12A, or Regnase-1) and a C-terminal truncation mutant (W543*) of this gene were inserted into pcDNA3.1(+) (V79020, Thermo Fisher Scientific) together with the N-terminal Myc-tag sequence. The S438L mutant *ZC3H12A* was created using the QuikChange Lightning Site-Directed Mutagenesis Kit (210519, Agilent). Plasmid DNA was transfected into HCT116 cells using Lipofectamine 2000 (11668500, Thermo Fisher Scientific) according to the manufacturer's protocol.

Degradation assay of ZC3H12A upon cytokine stimulation

HCT116 cells were stimulated with 50 ng ml^{-1} recombinant human IL-17A (570502, BioLegend) or 10 ng ml^{-1} IL-1β (201-LB-005, R&D Systems). Cells were lysed in the lysis buffer (20 mM Tris-HCl (pH = 7.4), 150 mM NaCl, 0.5% NP-40, cOmplete Mini (EDTA-free) (11836170001, Roche), 200 units per ml RNaseOUT (10777019, Thermo Fisher Scientific)). Protein samples were resolved on a 7.5% PAGE gel (E-T7.5L, ATTO) and transferred to Immun-Blot PVDF membranes (1620177, Bio-Rad). The following antibodies were used for immunoblot analysis: monoclonal anti-c-Myc (M4439, Sigma-Aldrich), monoclonal anti-beta-actin

(sc-47778 HRP, Santa Cruz) and anti-mouse IgG HRP (NA9310-1ML, GE Healthcare). Band intensities were analysed using Amersham Imager 600 Analysis Software (GE Healthcare). For the analysis of protein/mRNA ratio, total RNAs were isolated using TRIzol Reagent (15596018, Thermo Fisher Scientific) and reverse-transcribed using ReverTra Ace qPCR RT Master Mix with gDNA Remover (FSQ-301, TOYOBO) according to the manufacturer's protocol. The cDNA fragments were amplified with PowerUp SYBR Green Master Mix (A25742, Thermo Fisher Scientific). Primers are described in Supplementary Table 11.

Degradation assay of ZC3H12A under BTRC inhibition

HeLa cells were infected with CS-TRE-Flag-BTRCΔF-PRE-Ubc-tTA-I2G lentivirus, and treatment with doxycycline (Dox) induced the expression of the dominant-negative form of BTRC (BTRC(ΔF))[19]. After administration of 100 ng ml^{-1} Dox, cells were stimulated with 50 ng ml^{-1} recombinant human IL-17A (570502, BioLegend) or 10 ng ml^{-1} IL-1β (201-LB-005, R&D Systems). Cells were lysed and protein samples were prepared as described above. The following antibodies were used for immunoblot analysis: polyclonal anti-ZC3H12A (HPA032053, Atlas Antibodies), polyclonal anti-Flag (F7425, MERCK), monoclonal anti-beta-actin and anti-rabbit IgG HRP (NA9340V, GE Healthcare).

Analysis of the effect of *NFKBIZ* on growth of cancer cells

sgRNAs targeting *NFKBIZ* were designed, and knockout efficiency was confirmed by Guide-it Mutation Detection Kit (Z1448N, TAKARA). Five human CRC-derived cell lines (HT29, SW480, DLD1, HCT116 and LoVo) were infected with a lentivirus library generated from pL-CRISPR.EFS.GFP (Addgene 57818), containing SpCas9 and sgRNAs that target *NFKBIZ* (sgNFKBIZ #1–6) or non-targeting control sgRNAs (non-targeting #1–4) at a multiplicity of infection of 0.2 (Extended Data Fig. 10a). GFP-positive cells were sorted using a FACSAria IIIu (BD Biosciences) three days after infection. Half of the sorted cells were used for DNA extraction and the remaining half were cultured in 6-well plates. After four weeks of culture, 50 ng of genomic DNA extracted from each culture was subjected to PCR amplification of integrated sgRNAs, together with the DNA from pre-culture cells, for amplicon sequencing using NextSeq 550 (Illumina). The fold change (FC^{sgRNA}) of the frequency of sgRNA-containing reads between pre-culture ($Freq^{sgRNA/Pre}$) and post-culture ($Freq^{sgRNA/Post}$) samples was calculated and normalized for the mean value of control sgRNAs as follows:

$$FC^{sgRNA} = Freq^{sgRNA/Post}/Freq^{sgRNA/Pre}$$

$$\text{Normalized } FC^{sgRNA} = FC^{sgRNA}/\text{mean FC for control sgRNAs.}$$

Significant differences between sgRNAs that target *NFKBIZ* and non-targeting sgRNAs were calculated by two-sided Mann–Whitney *U* test. PCR primers and sgRNA sequences are provided in Supplementary Tables 11 and 12, respectively.

Animal experiments

All in vivo experiments in animals were approved by Kyoto University Animal Care and Use Committee. Genetically engineered C57BL/6 mice were used. *Vil1-cre* mice (004586) and *R26-LacZ* mice (003309) were obtained from The Jackson Laboratory. *Nfkbiz* flox mice[31] (RBRC06410) were obtained from Riken BRC. *Vil1-cre^{ER}* mice[32] were a gift from S. Robine.

Cellular response to inflammatory stimulation in colon organoids

Colon organoids were generated from *Nfkbiz$^{fl/fl}$Vil1-cre* (cKO) mice and *Nfkbiz$^{fl/fl}$* (control) mice and were maintained in basement membrane matrix (Matrigel) (354234, Corning) with 50% L-WRN conditioned medium (CM) as previously described[33]. Control organoids were stimulated with 50 ng ml^{-1} IL-1β (575102, BioLegend), 50 ng ml^{-1} IL-4 (574304, BioLegend), 50 ng ml^{-1} IL-13 (575904, BioLegend), 50 ng ml^{-1} IL-17A (421-ML-025, R&D Systems), 50 ng ml^{-1} IL-19 (2915-IL-025, R&D Systems), 50 ng ml^{-1} IL-20 (1204-ML-025, R&D Systems), 50 ng ml^{-1} IL-24 (7807-ML-010, R&D Systems), 50 ng ml^{-1} IL-31 (210-31-10UG, PeproTech), 50 ng ml^{-1} oncostatin M (762804, BioLegend), 50 ng ml^{-1} TNFα (575202, BioLegend), 50 ng ml^{-1} lymphotoxin α1/β2 (9968-LY-025/CF, R&D Systems), 10 μg ml^{-1} LPS (L6529-1MG, Sigma), 2.5 μM CpG oligodeoxynucleotides (ALX-746-001-C100, Enzo) and PBS (control) for 1 h or 24 h, followed by RNA extraction. The expression levels of *Nfkbiz* and *Gapdh* were measured by quantitative PCR. The expression levels of *Nfkbiz*, *Pigr* and *Gapdh* of organoids stimulated with IL-17A were also measured by quantitative PCR at 0, 1, 4 and 24 h after stimulation. For each gene, the expression level was calculated as a ratio relative to *Gapdh*. Primer sequences are described in Supplementary Table 11.

Competition assay of colon organoids from *Nfkbiz*$^{-/-}$ and *Nfkbiz*$^{+/+}$ mice

An equal number of colon organoids derived from *Nfkbiz$^{fl/fl}$Vil1-cre* (*Nfkbiz*$^{-/-}$) and *Nfkbiz$^{fl/fl}$* (*Nfkbiz*$^{+/+}$) mice were co-cultured in 50% L-WRN CM with or without 50 ng ml^{-1} IL-17A stimulation. Organoids were passaged in the ratio of 1:5 every 5 days. Genomic DNA was extracted at every passage, followed by measurement of *Nfkbiz*$^{-}$ and *Nfkbiz^{fl}* alleles using quantitative PCR. For each passage, the ratio of *Nfkbiz*$^{-}$ to *Nfkbiz^{fl}* alleles was calculated. Then, the fold change from the ratio at the start of culture was calculated for every passage. Primer sequences are described in Supplementary Table 11.

Clonal expansion of *Nfkbiz*-knockout colorectal epithelium in chronic DSS-induced colitis

R26$^{LacZ/WT}$Nfkbiz$^{fl/fl}$Vil1-cre^{ER} (cKO) mice and *R26$^{LacZ/WT}$Nfkbiz$^{WT/WT}$Vil1-cre^{ER}* (control) mice were co-housed after weaning. At 6 weeks of age, Cre-mediated recombination was induced by intraperitoneal injection of 2 mg of tamoxifen (T5648, SIGMA) twice a day for 2 consecutive days. At 8 to 9 weeks of age, experimental chronic colitis was induced by three cycles of 1% DSS (molecular weight = 36,000–50,000 Da; MP Biomedicals) (5 days DSS, followed by 16 days water). During three cycles of DSS, body weight was measured three times a week. Mice were killed before or after three cycles of DSS and colon lengths were measured. The colon was opened longitudinally and mounted, then stained by X-gal (9031, Takara) according to the manufacturer's protocol. Images of the colon were taken under a stereomicroscope and LacZ-positive areas from middle to distal colon were measured using ImageJ software. Histological scores of colitis in which epithelial damage and inflammatory-cell infiltration were scored from 0 to 12 as previously described[34] were counted on 5 randomly selected HPFs of H&E-stained slides in each mouse. The mean values of the scores were used for the analysis.

Transcriptome analysis of colorectal epithelial cells in DSS-induced colitis

Colorectal epithelial cells were obtained from rectum mucosa from 8-week-old *Nfkbiz$^{fl/fl}$Vil1-cre* (cKO) and *Nfkbiz$^{fl/fl}$* (control) mice that underwent 4 days of 1% DSS treatment. RNA was extracted from epithelial cells in distal colon by using the AllPrep DNA/RNA Micro Kit (Qiagen). Libraries for RNA sequencing were prepared using the NEBNext Ultra RNA Library Prep kit for Illumina (New England Biolabs) and subjected to sequencing using a HiSeq 2500 instrument (Illumina) with a standard 125-bp paired-end protocol.

The sequencing reads were aligned to the mouse reference genome (mm9) using STAR (v.2.5.3). Reads on each gene were counted with featureCounts (v.1.5.3) from the Subread package, and the edgeR package was used to identify the differentially expressed genes with a false discovery rate (FDR; *q* value) threshold of 0.05. The analysis was performed in genes that were expressed at >1 counts per million (CPM) in two or more samples, and generalized linear models were used to compare gene expression data. Gene set enrichment analysis (GSEA) (v.2.2.4) was used to determine the sets of genes that were differentially expressed in

Nfkbiz cKO mice compared with control mice. For GSEA, publicly available gene sets, including downregulated genes in *Il17rc*-knockout cells[35] and downregulated genes in *Nfkbiz*-knockout cells[13,36–38], were used.

Mouse model of cancer induced by AOM–DSS

Nfkbiz$^{fl/fl}$*Vil1-cre* (cKO) and *Nfkbiz*$^{fl/fl}$ (control) littermates were co-housed after weaning. At 8 weeks of age, 12 mg per gram body weight of AOM (A5486, SIGMA) was injected into the peritoneum and 3 cycles of 2% DSS were administered (5 days DSS, followed by 16 days water). After completion of 3 cycles of DSS, mice were killed and the colons were opened longitudinally. The size and number of tumours in the colon were measured. Induced tumours were evaluated histologically by H&E staining, β-catenin and Ki-67 immunohistochemistry as previously described[39]. Body weight, colon length and histological score of colitis were also measured as described above.

Statistical analysis

Statistical analyses were performed using R (v.3.5.1). All *P* values were calculated by two-sided analysis unless otherwise specified. The Student's *t*-test, paired *t*-test, Fisher's exact test or Mann–Whitney *U* test was used for group comparisons. Multiple testing was corrected based on the Benjamini–Hochberg method. The linearity of the number of mutations in single crypts ($n = 43$) from 22 control individuals with age was evaluated on the basis of the Pearson's correlation coefficient in a linear regression model that assumed zero intercept, because the number of somatic mutations among an exome region at 0 years of age is assumed to be nearly 0. By the direct observation of mutations without culturing cells, the estimated rate of mutation in the colorectal epithelium was 0.61 mutations per exome per year in control individuals, which is comparable to that reported previously[40]. For the analysis of the rate of mutation of single crypts from 18 patients with UC, we included one single crypt per subject that showed the maximum number of mutations in the same patients with UC and all crypts ($n = 43$) from 22 control individuals. The rate of mutation of single crypts after UC onset was estimated by a linear regression model that incorporated both age of UC onset and disease duration in which control individuals were treated as having a disease duration of 0, which showed significant improvement from a model that incorporated age of individuals alone (by ANOVA) (Supplementary Table 13). For comparison between the left and right side of the colon, the analyses were confined to those samples from patients for whom samples were taken from both the left and right side of the colon. In each patient, mutations or MCFs were averaged for multiple samples in the left and right side of the colon before performing a paired *t*-test. The significance of the difference of maximum MCF of isolated bulk-crypt samples between control individuals and patients with UC was tested by two-sided Mann–Whitney *U* test. For the linearity of maximum MCF in bulk-crypt samples with UC duration, a linear regression model without assuming zero intercept was applied, because a small number of mutations with high MCF can be observed as somatic mosaicism even in control individuals. The significance of the correlation of UC duration with maximum MCF was tested using an *F* test.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

All of the WES and RNA-seq data have been deposited in the European Genome-phenome Archive under accession numbers EGAS00001003801 and EGAS00001003802, respectively. Data for Figs. 1–4 and Extended Data Figs. 3–10 are available as Source Data. All other data are available from the corresponding author on reasonable request.

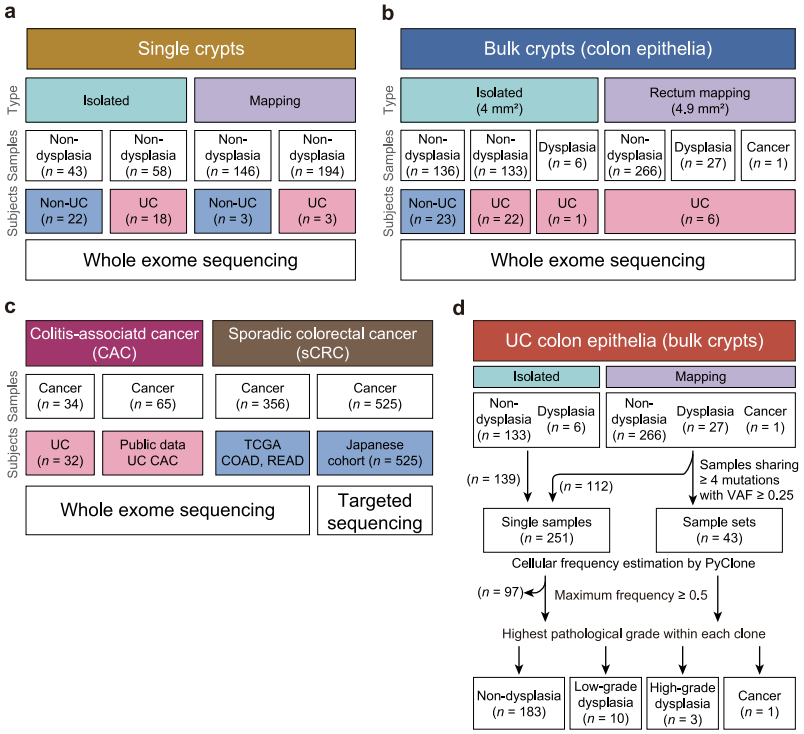
28. Riddell, R. H. et al. Dysplasia in inflammatory bowel disease: standardized classification with provisional clinical applications. *Hum. Pathol.* **14**, 931–968 (1983).
29. Yoshida, K. et al. Frequent pathway mutations of splicing machinery in myelodysplasia. *Nature* **478**, 64–69 (2011).
30. Shiraishi, Y. et al. An empirical Bayesian framework for somatic mutation detection from cancer genome sequencing data. *Nucleic Acids Res.* **41**, e89 (2013).
31. Okuma, A. et al. Enhanced apoptosis by disruption of the STAT3-IκB-ζ signaling pathway in epithelial cells induces Sjögren's syndrome-like autoimmune disease. *Immunity* **38**, 450–460 (2013).
32. el Marjou, F. et al. Tissue-specific and inducible Cre-mediated recombination in the gut epithelium. *Genesis* **39**, 186–193 (2004).
33. Miyoshi, H. & Stappenbeck, T. S. In vitro expansion and genetic modification of gastrointestinal stem cells in spheroid culture. *Nat. Protoc.* **8**, 2471–2482 (2013).
34. Katakura, K. et al. Toll-like receptor 9-induced type I IFN protects mice from experimental colitis. *J. Clin. Invest.* **115**, 695–702 (2005).
35. Chen, K. et al. IL-17 Receptor signaling in the lung epithelium is required for mucosal chemokine gradients and pulmonary host defense against *K. pneumoniae*. *Cell Host Microbe* **20**, 596–605 (2016).
36. Johansen, C. et al. IκBζ is a key driver in the development of psoriasis. *Proc. Natl Acad. Sci. USA* **112**, E5825–E5833 (2015).
37. Müller, A. et al. IκBζ is a key transcriptional regulator of IL-36-driven psoriasis-related gene expression in keratinocytes. *Proc. Natl Acad. Sci. USA* **115**, 10088–10093 (2018).
38. Hildebrand, D. G. et al. IκBζ is a transcriptional key regulator of CCL2/MCP-1. *J. Immunol.* **190**, 4812–4820 (2013).
39. Goto, N. et al. Distinct roles of Hes1 in normal stem cells and tumor stem-like cells of the intestine. *Cancer Res.* **77**, 3442–3454 (2017).
40. Blokzijl, F. et al. Tissue-specific mutation accumulation in human adult stem cells during life. *Nature* **538**, 260–264 (2016).

Acknowledgements This work was supported by the Japan Agency for Medical Research and Development (AMED); the Project for Development of Innovative Research on Cancer Therapeutics (JP19cm0106501 to S.Ogawa) and the Core Research for Evolutional Science and Technology (CREST) (JP19gm1110011 to S. Ogawa); the Ministry of Education, Culture, Sports, Science and Technology of Japan; the High Performance Computing Infrastructure System Research Project (hp160219, hp170227, hp180198 and hp190158 to S. Ogawa and S.M.) (this research used computational resources of the K computer provided by the RIKEN Advanced Institute for Computational Science through the HPCI System Research project); the Japan Society for the Promotion of Science (JSPS); Scientific Research on Innovative Areas (JP15H05909 to S. Ogawa and S.M.; JP15H05912 to S.M.; JP26115009 to M.S.) and KAKENHI (JP19H05656 to S. Ogawa; JP15H05707 to S.M.; JP17H04157 to H.S.; JP18H05278 to O.T.); the Ministry of Health, Labour and Welfare of Japan; Health and Labour Sciences Research Grants for research on intractable diseases (Investigation and Research for Intractable Inflammatory Bowel Disease) (H.N.); the Takeda Science Foundation (S. Ogawa, H.M. and T.Y.); the Uehara Memorial Foundation (H.S.); and the Foundation for Promotion of Cancer Research (H.S.). S. Ogawa is a recipient of the JSPS Core-to-Core Program A: Advanced Research Networks. We thank M. Nakazawa, M. Nakamura, T. Shirahari, A. Takatsu and A. Ryu for technical assistance; iLAC for sequencing support; the Centre for Anatomical, Pathological and Forensic Medical Research at Kyoto University Graduate School of Medicine for preparing microscope slides; and all staff at the Endoscopy Unit, Kyoto University Hospital for endoscopic sampling.

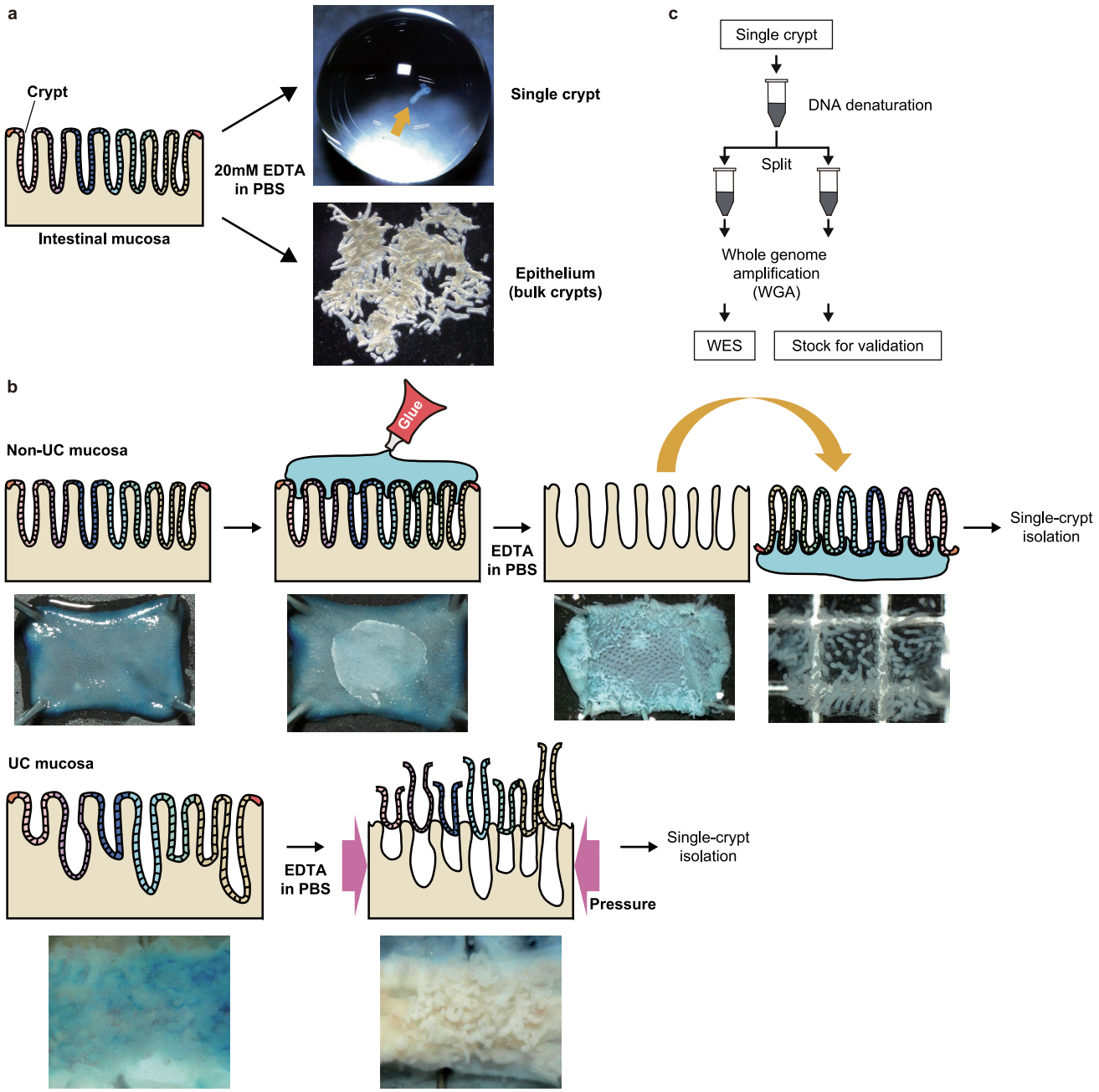
Author contributions N.K., K.Y., H. Marusawa, T.C., H.S. and S. Ogawa designed the study. M.U., K. Kawada, S.N., A. Yokoyama, S.Y., M.M., T. Horimatsu, S. Okamoto, H.O., Y. Sakai, H.I. and H.N. provided specimens. T.K., Y.T., T. Sakurai, H.H. and S.H. performed histological analysis. N.K., Y.I., Y.F. and S.K.K. performed sample preparation. M.S., E.S. and T. Sato performed sequencing. N.K., Y.O., Y. Shiozawa, T.Y. and Y.N. performed mutation calling, validated the results and analysed CNAs, mutational signatures and clonal dynamics. N.K., Y. Shiraishi, K.C., H.T. and S.M. performed bioinformatics analysis. N.K., K.A., T. Hirano, Y.K., Y.W., K. Kataoka, M.M.N., A. Yoda, H. Miyoshi, M.M.T., J.O., R.I., Y.U. and O.T. performed in vitro functional analysis. N.K., N.G., A.K., K. Kataoka and T.M. performed in vivo functional analysis. N.K., A. Yokoyama, Y.N., H. Makishima and S.Ogawa prepared the manuscript.

Competing interests The authors declare no competing interests.

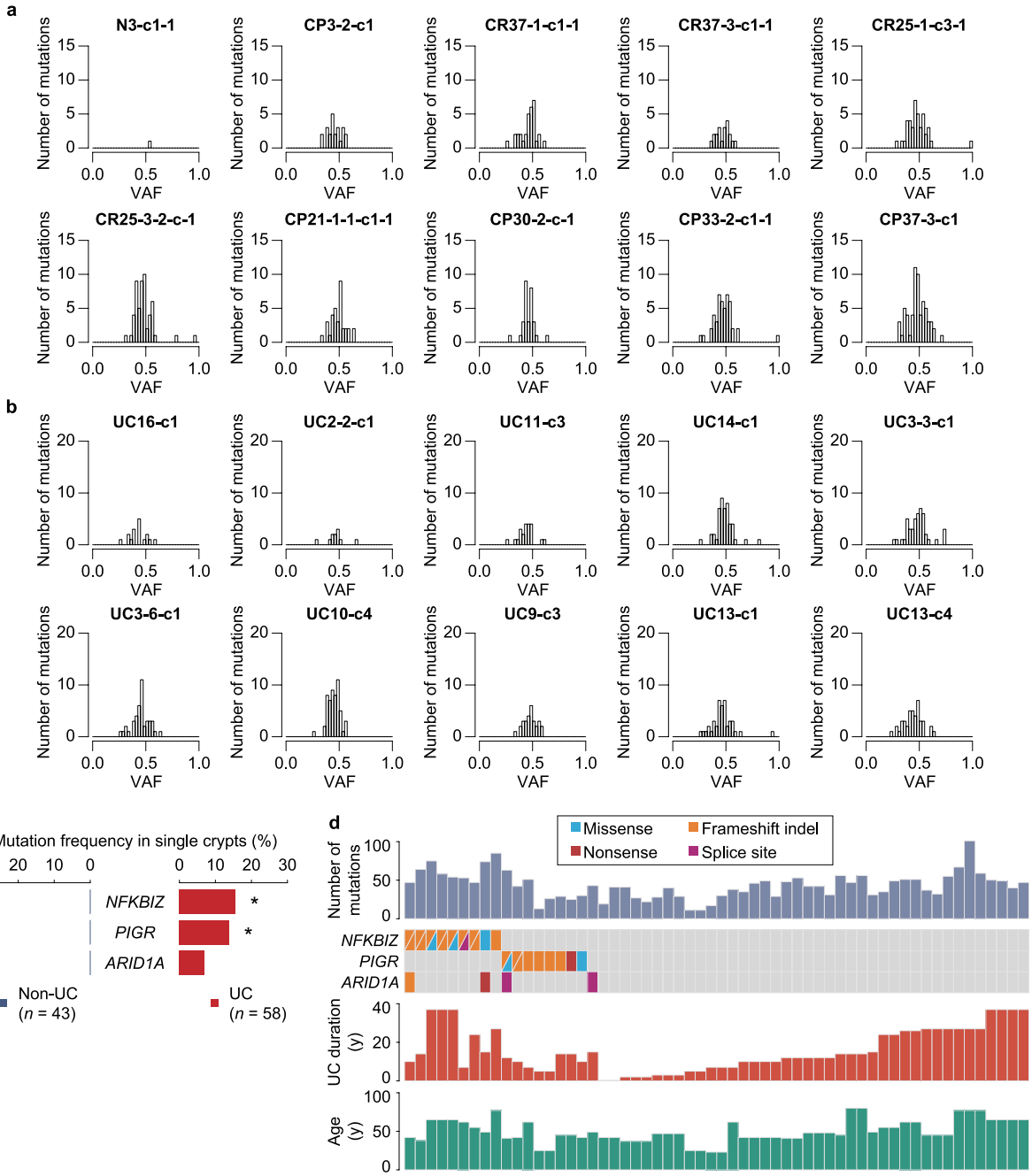
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1856-1.
Correspondence and requests for materials should be addressed to S.O.
Peer review information *Nature* thanks Judy Cho, Ramnik Xavier and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



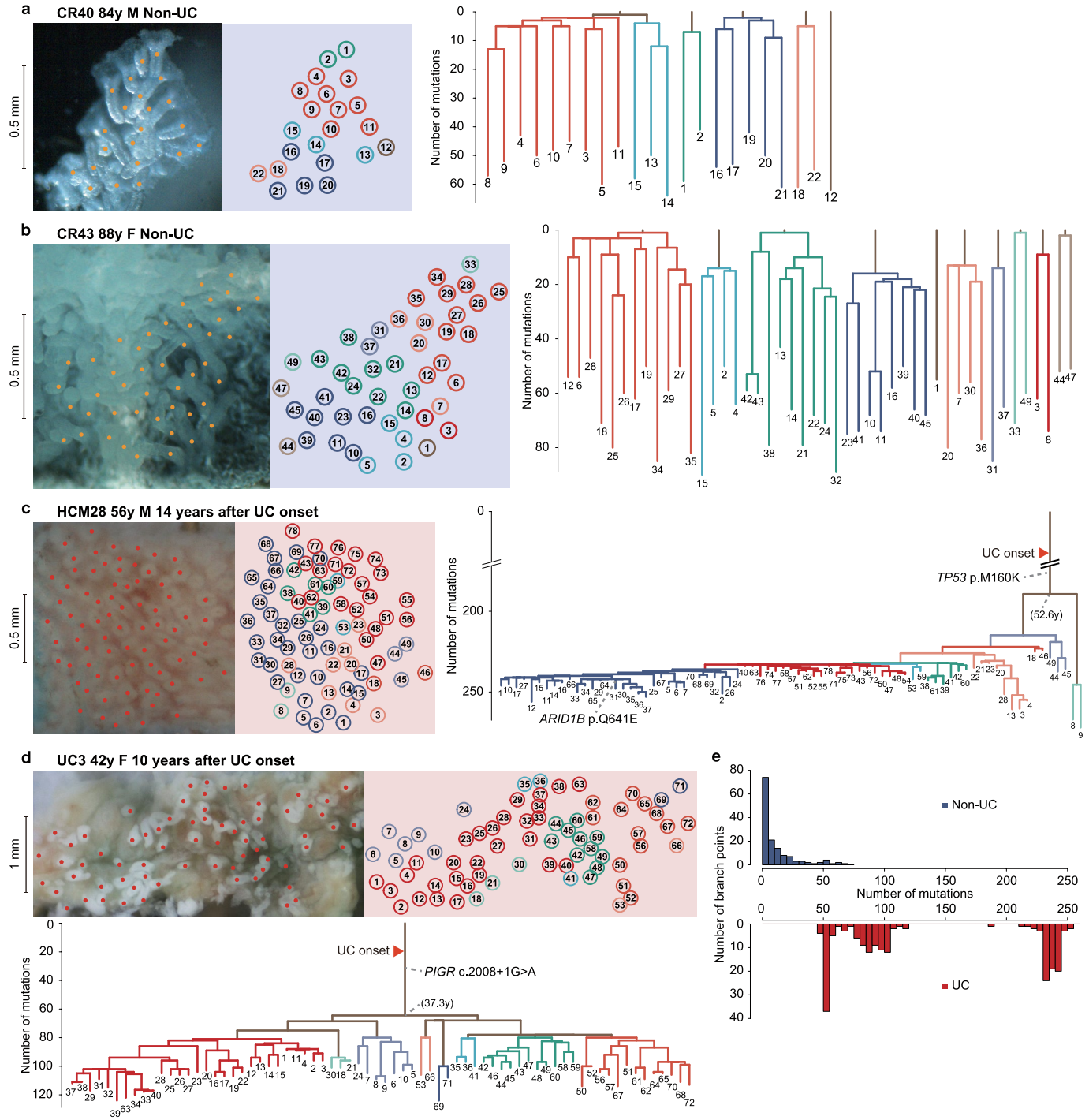
Extended Data Fig. 1 | Study design. a–**c**, Summary of sampling types, histology of samples, subjects and sequencing platforms for single crypts (**a**), bulk crypts (**b**) and colorectal cancers (**c**). COAD, TCGA Colon Adenocarcinoma data collection; READ, TCGA Rectum Adenocarcinoma data collection. **d**, Summary of independent clones as determined by PyClone analysis of bulk-crypt samples from patients with UC.



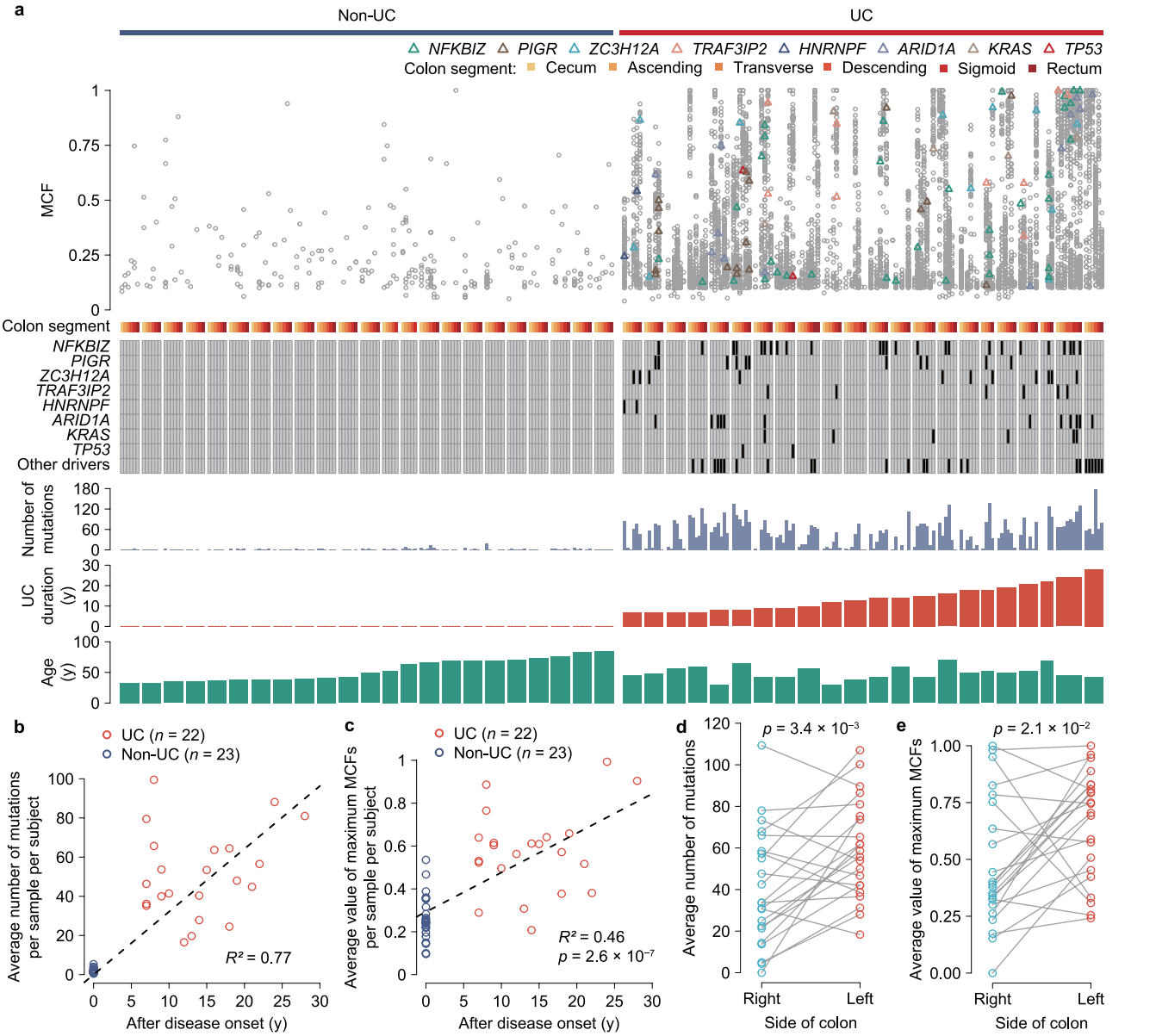
Extended Data Fig. 2 | Sampling methods. a, For preparation of epithelial samples, after treating intestinal mucosa with 20 mM EDTA in PBS at 4 °C for 20 min, the epithelium was dissociated from the lamina propria. Then, a single crypt or epithelium (for bulk crypts) was collected. **b**, For geographical tracking of clones in a cluster of single crypts in colorectal mucosa from control individuals, acrylic adhesive was put on the mucosal surface, followed by a treatment with 20 mM EDTA in PBS. After peeling off the epithelium upside down, single crypts that were attached to the adhesive were collected one by one under a stereomicroscope and their positions recorded (top). To isolate crypts from patients with UC, after colorectal mucosa was treated with 20 mM EDTA in PBS, a gentle pressure was applied to the whole mucosal tissue from both sides to push up crypts, and each single crypt was isolated in a manner similar to that described above. **c**, For single-crypt sequencing, genomic DNA was extracted from single crypts by direct cell lysis and heat denaturation and split into two aliquots, each of which was subjected to WGA using Φ29 DNA polymerase. One of the two WGA products was subjected to WES, and the other was stocked for validation.



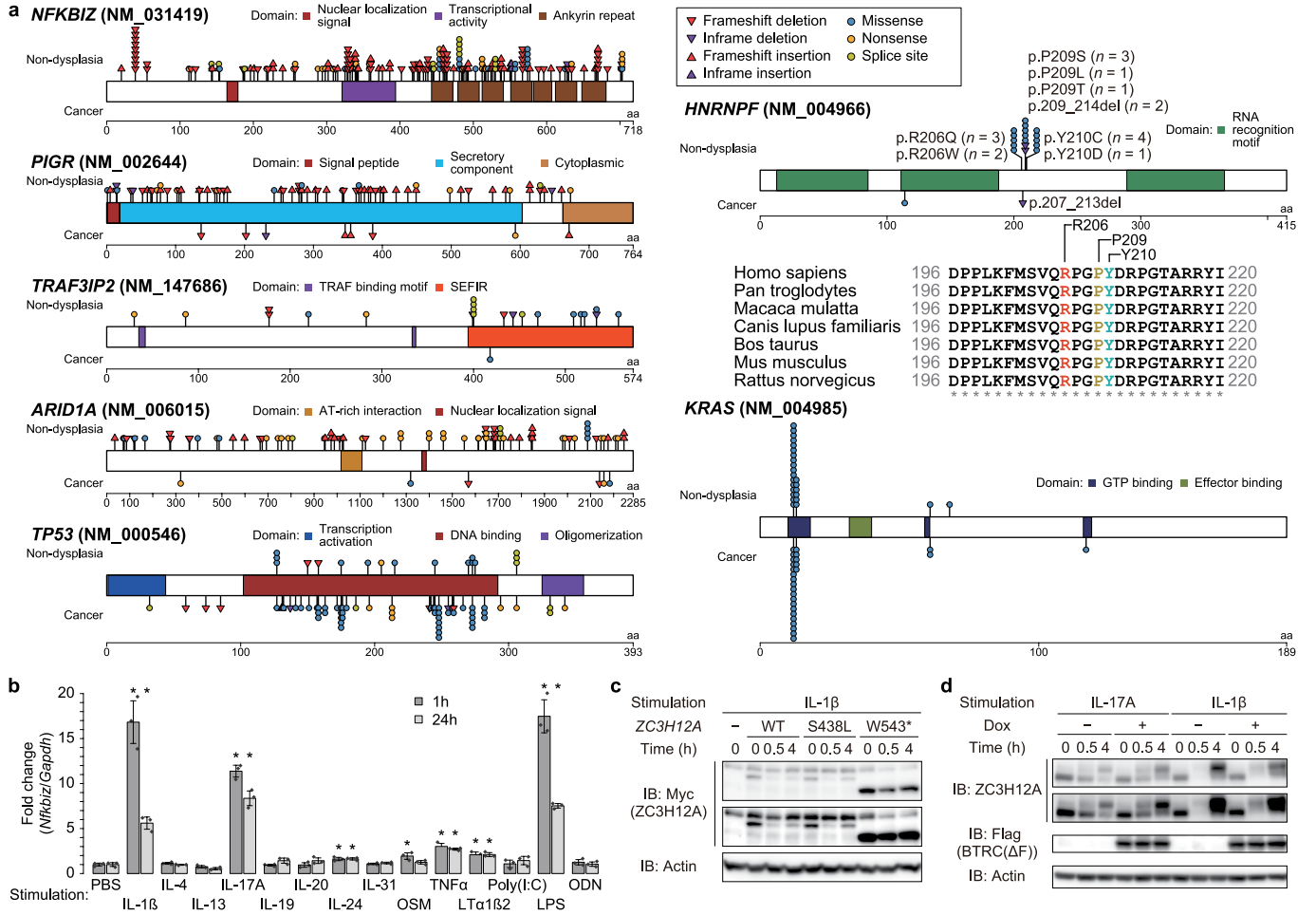
Extended Data Fig. 3 | Mutations detected in single crypts. **a**, **b**, Representative histograms of the VAFs of somatic mutations detected by WES in single crypts from control individuals (10 out of 43; **a**) and patients with UC (10 out of 58; **b**). **c**, Frequency of driver mutations among single crypts from control individuals ($n = 43$) and patients with UC ($n = 58$). Genes with a significant difference in mutation frequency between non-UC and UC crypts are indicated by asterisks (two-sided Fisher's exact test with Benjamini–Hochberg adjustment; $q < 0.1$). **d**, The landscape of driver mutations for single crypts ($n = 58$) from non-dysplastic UC samples. The number of mutations in each single crypt is shown above, and the disease duration and age of subjects are shown below.



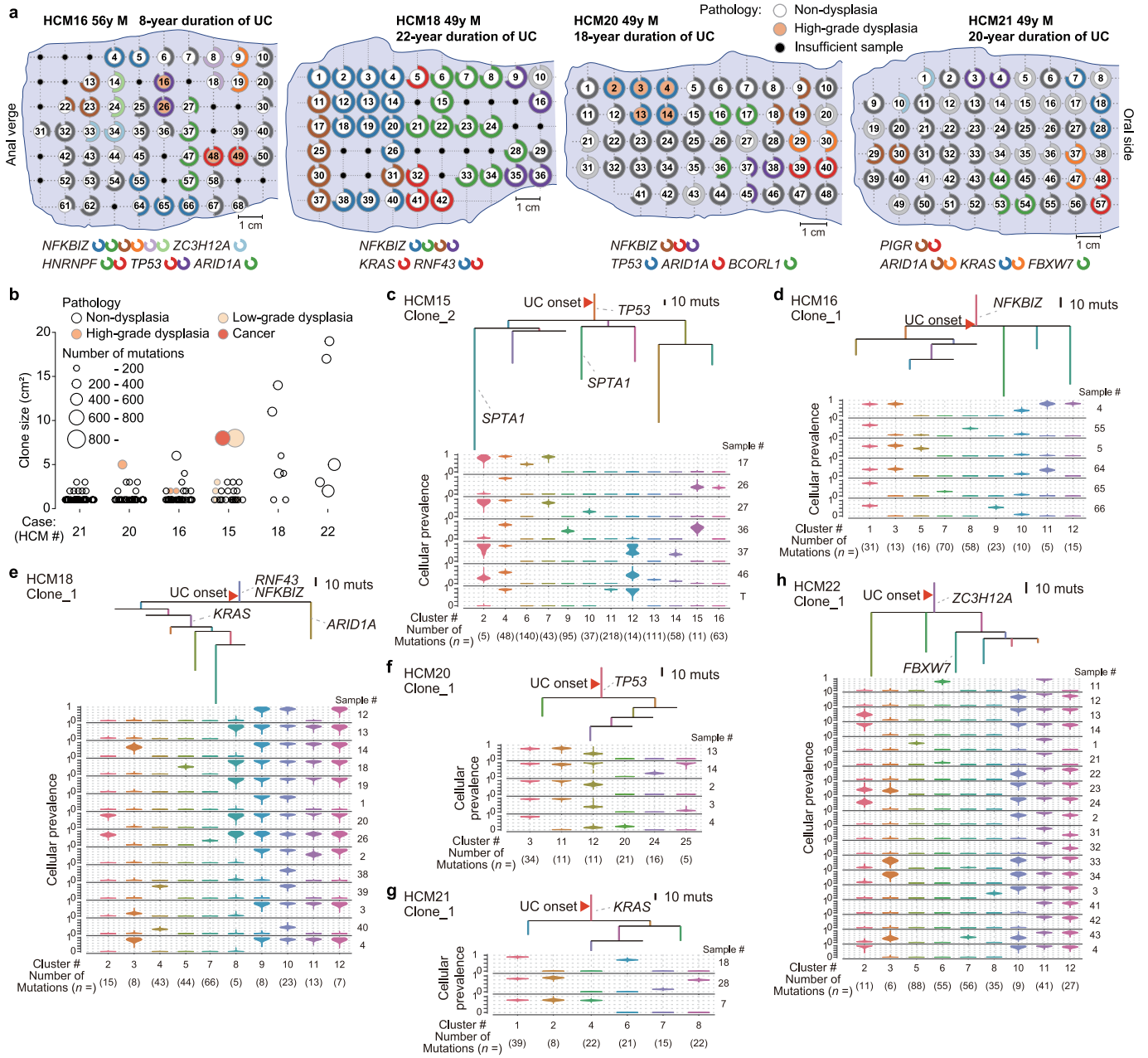
Extended Data Fig. 4 | Phylogenetic trees in the colorectal epithelium. **a**–**d**, Phylogenetic structure of single crypts clustered in an area of 0.25–4 mm^2 of epithelia from two control individuals (an 84-year-old man (CR40; **a**) and an 88-year-old woman (CR43; **b**)) and two patients with UC (a 56-year-old man (HCM28; **c**) and a 42-year-old woman (UC3; **d**)). Phylogenetic trees of clones (right or bottom) are projected onto the positions of crypts, as indicated in the images (left or top). The colours of the branches in the trees correspond to those of the circles for the positions of crypts. The estimated time of UC onset is indicated on the basis of the mutation rate in non-UC crypts (see Fig. 1a). The estimated age of the most-recent common ancestral crypt is also indicated for patients with UC (**c**, **d**). Branches that contain driver mutations are indicated. **e**, Histograms of the number of mutations (or height) of branch points in phylogenetic trees are plotted for control individuals ($n = 3$, top) and patients with UC ($n = 3$, bottom).



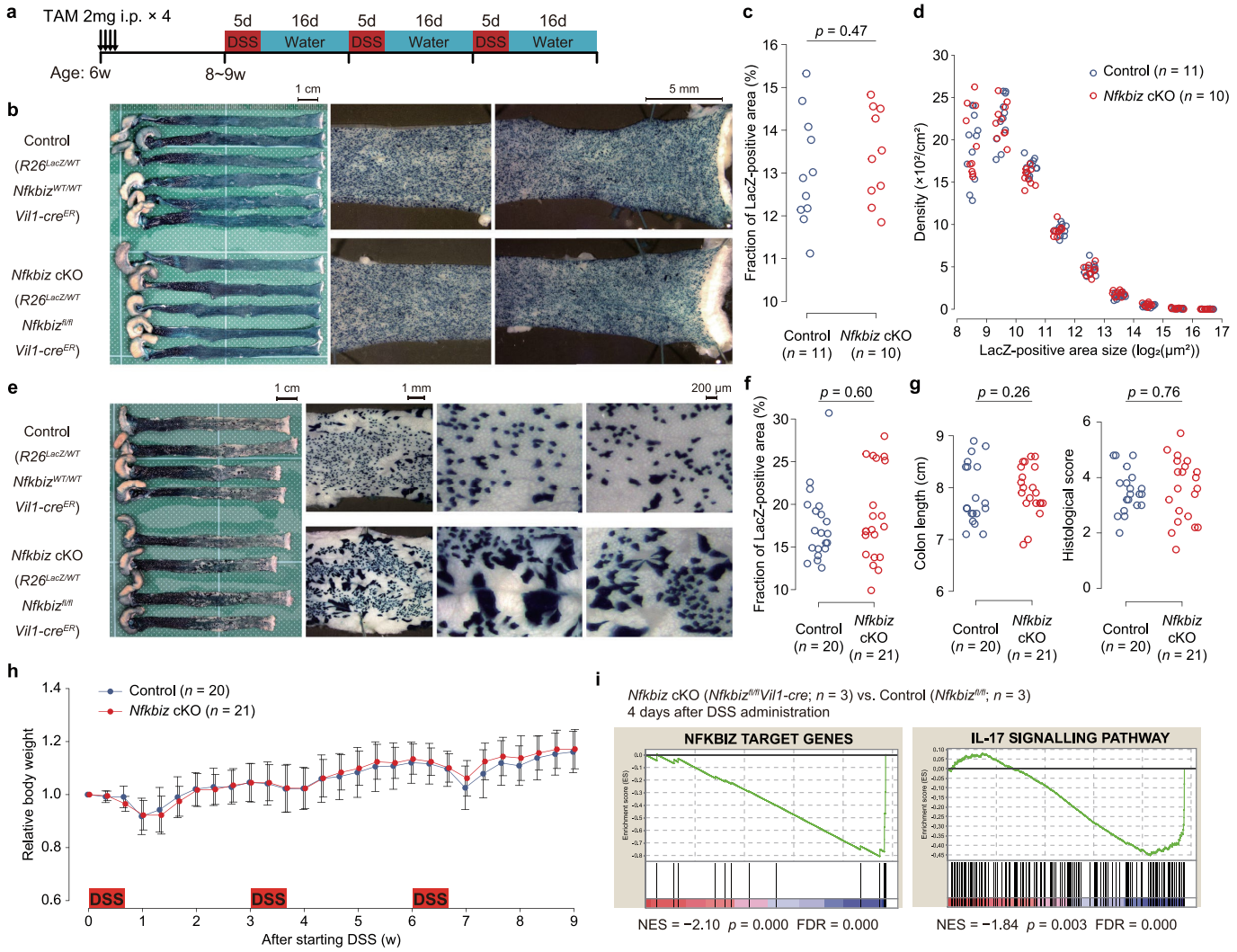
Extended Data Fig. 5 | Mutations detected in colorectal epithelium from control individuals and patients with UC. a, MCF of mutations that were detected with WES in isolated bulk-crypt samples from non-UC (n = 136) and UC (n = 133) colorectal epithelia. Samples belonging to each subject are separated by blank columns. MCFs of driver mutations are indicated by coloured triangles, and disease state by colour, as shown at the top. The colon segment that was sampled, mutation status of driver genes, number of mutations, duration of UC and age of subjects are shown at the bottom. **b**, **c**, The average number of mutations (**b**) and average value of maximum MCFs (**c**) detected by WES of bulk-crypt samples within each individual is plotted against the duration of UC. Control individuals are plotted at zero on the x axis. Regression lines with (**b**) or without (**c**) assuming zero intercept are shown with R^2 values. The calculated P value in **c** represents the significance of the correlation between the duration of UC and the average value of maximum MCFs (F test). **d**, **e**, The average number of mutations (**d**) and average value of maximum MCFs (**e**) of bulk-crypt samples from caecum to transverse colon (right side) and descending colon to rectum (left side) within the same individual is plotted for patients with UC (n = 22). Two-sided paired t-test.



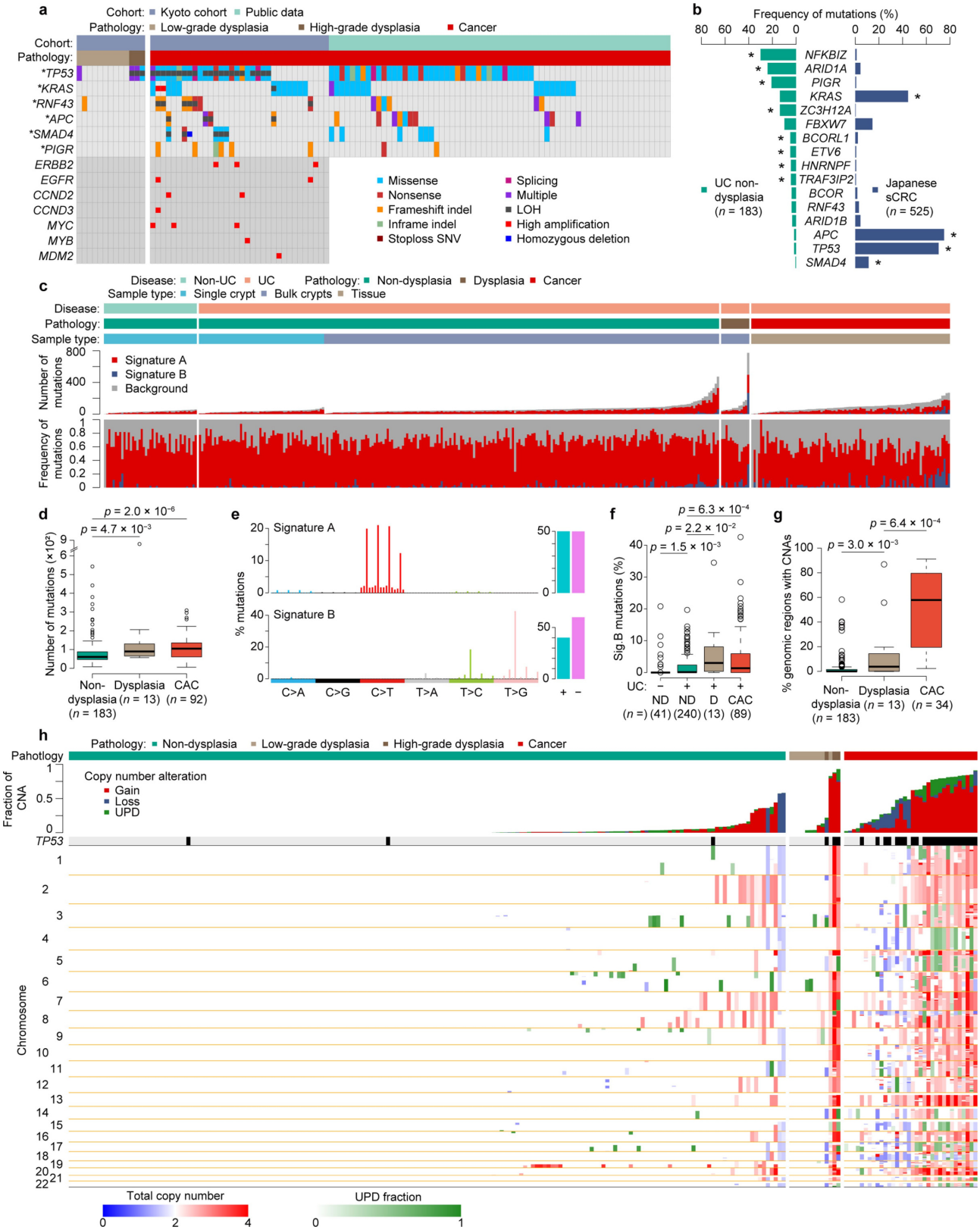
Extended Data Fig. 6 | Functional analysis of driver genes in UC mucosa. **a**, Distribution of driver gene mutations in non-dysplastic UC samples (58 isolated single crypts and 399 bulk crypts, top) and CAC samples ($n = 99$, bottom). Amino acid sequences of different species around hotspot mutations of *HNRNPF* are shown. Completely conserved amino acids across all species are indicated by asterisks. **b**, Expression levels of *Nfkbiz* in colon epithelial organoids ($n = 3$ biological replicates) treated as indicated for 1 or 24 h. Data are normalized by the levels of PBS (control) and shown as mean ± s.d. Asterisks indicate $P < 0.05$ between the indicated treatment and treatment with PBS (two-sided Student's *t*-test). LTα1β2, lymphotoxin α1/β2; ODN, CpG oligodeoxynucleotides; OSM, oncostatin M. **c**, Immunoblot analysis of HCT116 cells that express empty vector, Myc-tagged wild-type (WT) *ZC3H12A* or the indicated *ZC3H12A* mutants after stimulation with IL-1β for the indicated times, using anti-Myc and β-actin antibodies. **d**, Immunoblot analysis of ZC3H12A, Flag-tagged dominant-negative BTRC(ΔF) and β-actin in HeLa cells that express doxycycline-inducible BTRC(ΔF) after stimulation with IL-17A or IL-1β for the indicated times with or without doxycycline (Dox). Data are representative of two independent experiments. For gel source data, see Supplementary Fig. 1.



Extended Data Fig. 7 | Clonal expansion in UC rectums. a, Expansion of clones within the rectum from the remaining four out of the six analysed patients with UC (two are presented in Fig. 2d). All of the representations follow those in Fig. 2d. Driver mutations including those found in the two samples in Fig. 2d are summarized in Supplementary Table 9. **b**, Area occupied by each clone and total number of mutations estimated by PyClone for each clone detected in the UC rectums. Colours indicate the highest pathological grade within each clone. **c**–**h**, Phylogenetic trees of representative clones (6 out of 41) that were found in two or more rectum samples surgically obtained from six patients with UC. Imputed clonal structures are depicted in violin plots at the bottom, in which the estimated cellular prevalence of imputed mutational clusters (*x* axis) are shown as the distribution of posterior probabilities (width of the violin plots) calculated according to a PyClone model. Colours correspond to those in the associated branches of trees. The estimated time of UC onset is indicated on the basis of the mutation rate in non-UC crypts (Fig. 1a). Branches that contain driver mutations are also indicated.

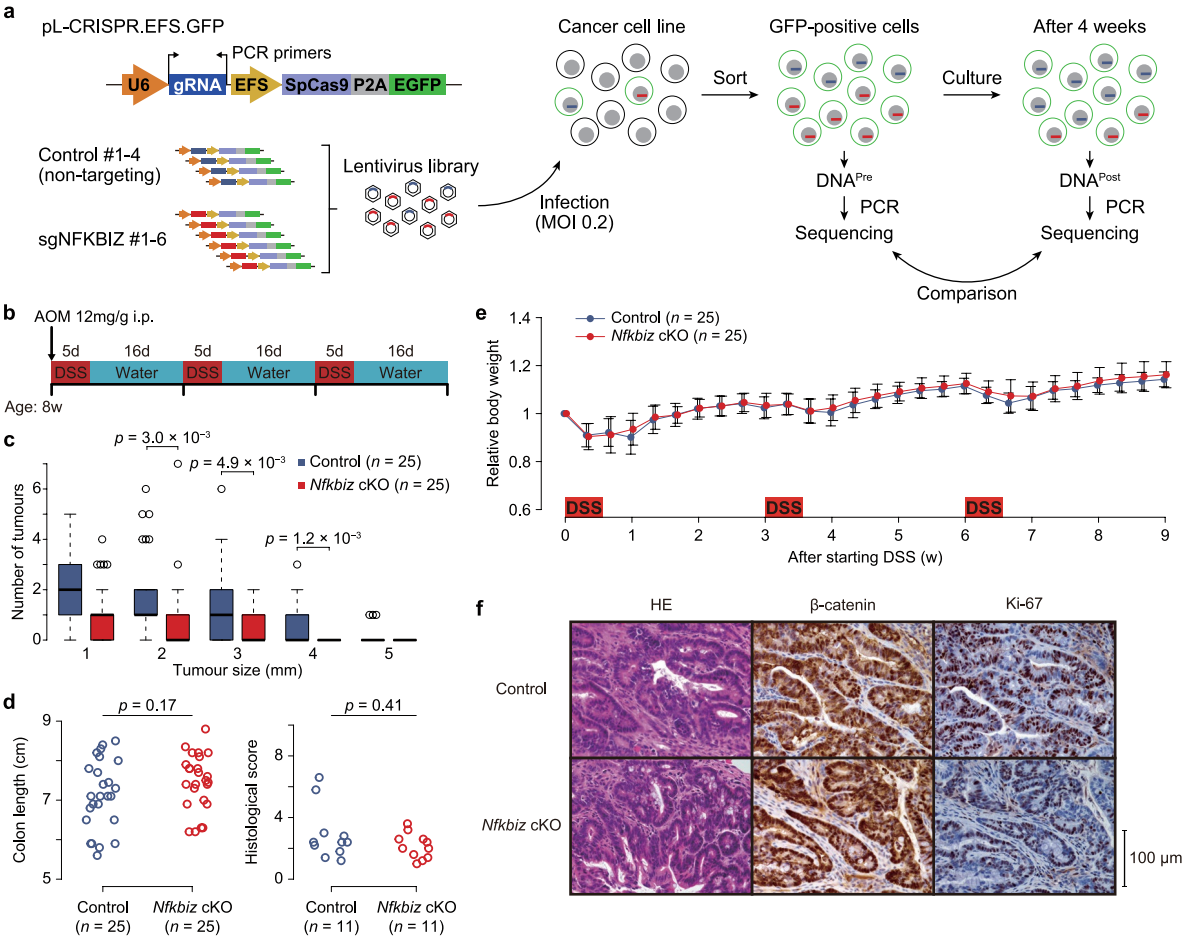


Extended Data Fig. 8 | Effect of *Nfkbiz* on clonal expansion in a mouse model of chronic DSS-induced colitis. a, Experimental schedule for induction of colitis in $R26^{LacZ/WT}Nfkbiz^{WT/WT}Vil1\text{-}cre^{ER}$ (control) and $R26^{LacZ/WT}Nfkbiz^{fl/fl}Vil1\text{-}cre^{ER}$ (*Nfkbiz* cKO) mice. Cre-mediated recombination was induced by administration of tamoxifen (TAM), followed by induction of colitis by treatment with DSS. **b**, **e**, Images of whole-mount X-gal staining of the large intestine from control (top) and *Nfkbiz* cKO (bottom) mice after tamoxifen administration without (**b**) or with (**e**) subsequent treatment with DSS. **c**, **f**, Fraction of LacZ-positive areas in the middle to distal colon before (**c**) or after (**f**) induction of colitis with DSS according to genotypes. Two-sided Mann–Whitney *U* test. **d**, Density of LacZ-positive areas (number per cm²) having indicated sizes in control and *Nfkbiz* cKO mice before induction of colitis with DSS. **g**, Colon length (left) and histological score (right) after chronic DSS-induced colitis for the evaluation of severity of colitis. Two-sided Mann–Whitney *U* test. **h**, Relative change in weight of *Nfkbiz* cKO and control mice during three courses of DSS-induced colitis. Data are shown as mean ± s.d. **i**, Significant enrichment of NFKBIZ target genes (left) and genes associated with the IL-17 signalling pathway (right) among the downregulated genes in rectum epithelium from *Nfkbiz* cKO mice compared with control mice four days after the start of treatment with DSS. Data were generated from $n = 3$ biologically independent mice for both genotypes. Two-sided nominal *P* values were calculated by GSEA.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Mutational profiles and CNAs in control individuals and patients with UC. a, Landscape of driver mutations in CAC samples ($n = 99$) and dysplastic UC samples ($n = 13$). Genes with asterisks are drivers in CACs (Methods). CNAs of known oncogenes are depicted only for the 34 CACs that were newly analysed in this study. **b**, Frequency of driver mutations among clones identified in non-dysplastic samples from patients with UC ($n = 183$) versus sCRC samples from the Japanese cohort ($n = 525$). Asterisks indicate $q < 0.1$ (two-sided Fisher's exact test with Benjamini–Hochberg adjustment). **c**, Number (middle) and relative frequency (bottom) of mutations allocated to signatures A and B in samples of single crypts from control individuals ($n = 43$) and non-dysplastic ($n = 241$; 58 single crypts and 183 clones), dysplastic ($n = 13$ clones) and non-hypermutated CAC ($n = 92$ samples) from patients with UC. Information about disease state, pathology and sample type is indicated at the top. **d**, Comparison of the number of mutations detected by WES in clones identified in non-dysplastic samples, dysplastic samples and CAC samples without hypermutation from patients with UC. Two-sided Mann–Whitney *U* test. **e**, Mutational signatures of colorectal epithelium from control individuals and patients with UC. **f**, Fraction of signature B mutations in single crypts from control individuals, and clones identified in non-dysplastic (ND), dysplastic (D) and non-hypermutated CAC samples containing ≥ 10 mutations from patients with UC. Two-sided Mann–Whitney *U* test. **g**, Box plots of the fractions of genomic regions that contain CNAs in clones identified in non-dysplastic ($n = 183$), dysplastic ($n = 13$) and CAC ($n = 34$) samples from patients with UC. Two-sided Mann–Whitney *U* test. **h**, Bottom, colour-gradient maps of CNAs and UPDs as detected by WES data are shown for clones identified in non-dysplastic ($n = 183$), dysplastic ($n = 13$) and CAC ($n = 34$) samples from patients with UC. Top, pathological grade, fractions of genomes that exhibit CNAs and UPDs and *TP53* mutation status are also shown. Box plots as in Fig. 4.



Extended Data Fig. 10 | Effect of *NFKBIZ* knockout in CRC cell lines and an AOM–DSS-induced colorectal tumour model. a, Experimental strategy for enrichment analysis using CRISPR–Cas9 (Methods). MOI, multiplicity of infection. **b**, Experimental schedule for induction of CAC in *Nfkbiz*$^{fl/fl}$*Vil1-cre* (*Nfkbiz* cKO) and *Nfkbiz*$^{fl/fl}$ (control) mice using AOM followed by three cycles of DSS. **c**, Box plots of tumour numbers relative to tumour size (maximum diameter). Two-sided Mann–Whitney *U* test. **d**, Colon length (left) and histological score (right) for the evaluation of inflammation in the AOM–DSS-induced tumour model. No significant difference was observed between *Nfkbiz* cKO and control mice (two-sided Mann–Whitney *U* test). **e**, Relative change in weight of mice during three courses of treatment with DSS in the AOM–DSS-induced tumour model. Data are shown as mean ± s.d. **f**, Histological images of haematoxylin and eosin (H&E) staining (left), and immunohistochemistry for β-catenin (middle) and Ki-67 (right). Tumours in *Nfkbiz* cKO and control mice were indistinguishable in terms of histology and immunochemistry. Representative images are shown from *n* = 3 biologically independent samples for both genotypes. Box plots as in Fig. 4.

nature research

Corresponding author(s): Seishi Ogawa

Last updated by author(s): Nov 4, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	We enrolled a total of 647 subjects who underwent surgery or endoscopy of the lower gastrointestinal tract at Kyoto University Hospital, Hyogo College of Medicine Hospital, and Cancer Institute Hospital. Among these, 76 had been diagnosed as ulcerative colitis and the remaining 571 were diagnosed as other disease such as colorectal polyp or cancer without colitis, or were turned out to be healthy individuals. This study was approved by the ethical committees at Kyoto University, Hyogo College of Medicine, and Cancer Institute Hospital. Informed consent was obtained from all participants. As for external data sets, whole exome sequencing (WES) data of paired tumor/germline control samples from CRC (n=356) patients were downloaded from the TCGA Data Portal (https://portal.gdc.cancer.gov/). Four previously published mutation lists of WES obtained from 65 colitis-accociated cancer in UC individuals were also analysed.
Data analysis	Detailed descriptions of the software and analysis have been provided in Online Methods. Sequencing data was processed using our in-house pipeline Genomon2. External bam files were converted to fastq format using biobambam2. Copy number analysis using sequencing data was performed by in-house pipeline CNACS. Significantly mutated genes were identified using dNdScv and MutSigCV. Mutational signature was evaluated using pmsignature and deconstructSigs. Clonal architecture was estimated by PyClone (0.13.0). Phylogenetic analysis was performed by using MEGA X (v10.0.5). Transcriptome analysis was performed by STAR (v2.5.3.), featureCounts (v1.5.3), edgeR (v3.24.3), and GSEA (v2.2.4). Statistical analyses were performed using R (3.5.1). Images were analyzed by using ImageJ (1.52a). List of programs and softwares: Genomon2 pipeline: version 2.6.0 (https://genomon.readthedocs.io/ja/latest) - Burrows-Wheeler Aligner: version 0.7.10 (https://sourceforge.net/projects/bio-bwa/) - picard-tools: version 1.39 (http://picard.sourceforge.net/) - biobambam2: version 2.0.85 (https://www.sanger.ac.uk/science/tools/biobambam) - GenomonMutationFilter: (https://github.com/Genomon-Project/GenomonMutationFilter) - EBCall: version 2 (https://github.com/friend1ws/EBCall) Integrative Genomics Viewer (IGV): version 2.4.6 (http://software.broadinstitute.org/software/igv/)

MutSigCV: version 1.4. (http://software.broadinstitute.org/cancer/software/genepattern/modules/docs/MutSigCV)
dNdScv: version 0.0.1 (https://github.com/im3sanger/dndscv)
CNACS: (http://plaza.umin.ac.jp/kyoto_tumorpatho/CNACS.html)
pmsignature: version 0.3.0 (https://github.com/friend1ws/pmsignature)
deconstructSigs: version 1.8.0 (https://github.com/raerose01/deconstructSigs)
PyClone: version 0.13.0 (https://github.com/aroth85/pyclone)
MEGA X: version 10.0.5 (https://www.megasoftware.net/home)
STAR: v2.5.3 (https://github.com/alexdobin/STAR)
featureCounts: v1.5.3 (http://bioinf.wehi.edu.au/featureCounts/)
edgeR: 3.24.3 (https://bioconductor.org/packages/release/bioc/html/edgeR.html)
GSEA: v2.2.4 (http://software.broadinstitute.org/gsea/index.jsp)
R: v.3.5.1 (https://cran.r-project.org/)
ImageJ: v1.52a (https://imagej.nih.gov/ij/index.html)

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

All the WES and RNA-seq data have been deposited in the European Genome-Phenome Archive (http://www.ebi.ac.uk/ega/) under accession numbers EGAS00001003801 and EGAS00001003802, respectively.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	No statistical methods were used to determine sample size since this is an exploratory study. We enrolled as many patients as possible who provided consent for our study during the enrollment period between 1/Oct/2014 and 31/Mar/2019. Cryo-preserved colorectal cancer samples were provided from Cancer Institute Hospital. In total, we collected 1568 samples from 76 UC patients and 571 healthy or non-UC individuals. The sample size of our study was sufficient to confirm novel findings with statistical significance. As for external data sets, 356 sporadic CRCs (sCRCs) and 65 colitis-associated cancers were also analysed.
Data exclusions	For the analysis of TCGA sCRC samples, we excluded whole-genome amplified samples to avoid artefactual mutation callings. We also excluded hypermutated TCGA sCRCs from the analysis of comparison of driver genes between UC non-dysplasia and TCGA sCRCs because high background mutation rate of the tumours might make the difference difficult to detect. We also enrolled only non-hypermutated sCRCs from the Japanese cohort for the same reason.
Replication	Three or more biological replicates were generated for every assay, except for immunoblot experiments where two biological replicates were generated and the same results were confirmed among two replicates. We tested significant difference for every assay as described in the figure legends and Methods.
Randomization	Not applicable since this is a case-series study which was therefore not planned to detect any difference in effects between the cohorts with and without intervention.
Blinding	Two expert pathologists, blinded to clinical information, reviewed all the specimens obtained from UC patients. Researchers were blinded to genotype information at measurement of each animal experiment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☐	☒ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used

Monoclonal anti-c-Myc (supplier name, Sigma-Aldrich; catalogue number, M4439; clone name, 9E10; lot number, 026M4825V; dilution, 1/2000)
Anti-mouse IgG HRP (supplier name, GE Healthcare; catalogue number, NA9310-1ML; lot number, 16919626; dilution, 1/5000)
Polyclonal anti-ZC3H12A (supplier name, Atlas Antibodies; catalogue number, HPA032053; lot number, R32479; dilution, 1/500)
Polyclonal anti-Flag (supplier name, Sigma-Aldrich; catalogue number, F7425; lot number, 078M4886V; dilution, 1/2000)
Monoclonal anti-beta-Actin (supplier name, Santa Cruz; catalogue number, sc-47778 HRP; clone name, C4; lot number, K1418; dilution, 1/1000)
Anti-rabbit IgG HRP (supplier name, GE healthcare; catalogue number, NA9340V; lot number, 16926347; dilution, 1/5000)

Validation

Monoclonal anti-c-Myc (supplier name, Sigma-Aldrich; catalogue number, M4439) – immunoblotting – human cells; Validation: Manufacturer - https://www.sigmaaldrich.com/catalog/product/sigma/m4439
Polyclonal anti-ZC3H12A (supplier name, Atlas Antibodies; catalogue number, HPA032053) – immunoblotting – human cells; Validation: Manufacturer - https://www.atlasantibodies.com/products/antibodies/primary-antibodies/triple-a-polyclonals/zc3h12a-antibody-hpa032053/
Polyclonal anti-Flag (supplier name, MERCK; catalogue number, F7425) – immunoblotting – human cells; Validation: Manufacturer - https://www.sigmaaldrich.com/catalog/product/sigma/f7425
Monoclonal anti-beta-Actin (supplier name, Santa Cruz; catalogue number, sc-47778 HRP) – immunoblotting – human cells; Validation: Manufacturer - https://datasheets.scbt.com/sc-47778.pdf

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)

HT29, SW480, DLD1, HCT116, LoVo, and HeLa cells were obtained from ATCC.

Authentication

None of the cell lines were authenticated.

Mycoplasma contamination

All cell lines used for this study were tested negative for mycoplasma using Myco Alert Mycoplasma Detection Kit (Lonza).

Commonly misidentified lines
(See ICLAC register)

No commonly misidentified cell lines were used in this study.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals

VilCre mice (004586) and R26-LacZ mice (003309) were obtained from The Jackson Laboratory. Nfkbiz flox mice (RBRC06410) were obtained from Riken BRC. VilCreER mice were kindly gifted by Sylvie Robine. All mice used in this study were C57BL/6 strains.
In the analysis of clonal expansion in chronic DSS colitis, we analysed 11 R26-LacZ/wt;Nfkbiz wt/wt;VilCreER (control) mice (5 male and 6 female) and 10 R26-LacZ/wt;Nfkbiz fl/fl;VilCreER (Nfkbiz cKO) mice (5 male and 5 female) for the pre-colitis assessment. We also analysed 20 control mice (13 male and 7 female) and 21 Nfkbiz cKO mice (14 male and 7 female) for the post-colitis assessment. All mice were injected with tamoxifen at 6 weeks of age.
In the transcriptome analysis of DSS-induced colitis, 3 male Nfkbiz fl/fl (control) mice and 3 male Nfkbiz fl/fl;VilCre (Nfkbiz cKO) mice were analysed at 8 weeks of age.
In the AOM-DSS tumour model, we started experiments with 25 Nfkbiz fl/fl (control) mice (18 male and 7 female) and 25 Nfkbiz fl/fl;VilCre (Nfkbiz cKO) mice (18 male and 7 female) at 8 weeks of age.

Wild animals

This study did not involve wild animals

Field-collected samples

This study did not involve field-collected samples

Ethics oversight

All in vivo experiments in animals were approved by Kyoto University Animal Care and Use Committee.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Participants were recruited through at Kyoto University, Hyogo College of Medicine, and Cancer Institute Hospital. We enrolled a total of 647 patients who underwent surgery or endoscopy of the lower gastrointestinal tract at Kyoto University Hospital, Hyogo College of Medicine Hospital, and Cancer Institute Hospital. Among these, 76 had been diagnosed as ulcerative colitis (UC) and the remaining 571 (healthy or non-UC) were diagnosed as other disease such as colorectal polyp or cancer without associated colitis or turned out to be healthy individuals. The detailed characteristics of cases were provided in Supplementary Table 1.
Recruitment	Patients who underwent surgery or endoscopy of the lower gastrointestinal tract between 1/Oct/2014 and 31/Mar/2019 at Kyoto University and Hyogo College of Medicine were recruited without any selection. Cryo-preserved non-hypermutated sCRC samples between 1/Dec/2006 and 31/May/2016 were provided from Cancer Institute Hospital without any selection.
Ethics oversight	This study was approved by the ethical committees at Kyoto University, Hyogo College of Medicine, and Cancer Institute Hospital. Informed consent was obtained from all participants.

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Clinical data

Policy information about clinical studies

All manuscripts should comply with the ICMJE guidelines for publication of clinical research and a completed CONSORT checklist must be included with all submissions.

Clinical trial registration	Not applicable. This study is not a clinical trial.
Study protocol	Not applicable. This study is not a clinical trial.
Data collection	Not applicable. This study is not a clinical trial.
Outcomes	Not applicable. This study is not a clinical trial.

Article

HBO1 is required for the maintenance of leukaemia stem cells

https://doi.org/10.1038/s41586-019-1835-6

Received: 5 September 2018

Accepted: 12 November 2019

Published online: 11 December 2019

Laura MacPherson[1,2], Juliana Anokye[1], Miriam M. Yeung[1], Enid Y. N. Lam[1,2], Yih-Chih Chan[1,2], Chen-Fang Weng[1], Paul Yeh[1,2], Kathy Knezevic[1,2], Miriam S. Butler[1,2], Annabelle Hoegl[3], Kah-Lok Chan[1,2], Marian L. Burr[1,2], Linden J. Gearing[4,5], Tracy Willson[4,5], Joy Liu[4], Jarny Choi[4,5], Yuqing Yang[4,5], Rebecca A. Bilardi[4,5], Hendrik Falk[4,5,6], Nghi Nguyen[7], Paul A. Stupple[6,7], Thomas S. Peat[6,8], Ming Zhang[4,5,6], Melanie de Silva[4,5,6], Catalina Carrasco-Pozo[6,9], Vicky M. Avery[6,9], Poh Sim Khoo[6,10], Olan Dolezal[6,8], Matthew L. Dennis[6,8], Stewart Nuttall[6,8], Regina Surjadi[6,8], Janet Newman[6,8], Bin Ren[6,8], David J. Leaver[7], Yuxin Sun[7], Jonathan B. Baell[7,11], Oliver Dovey[12], George S. Vassiliou[12,13], Florian Grebien[14], Sarah-Jane Dawson[1,2,15], Ian P. Street[4,5,6], Brendon J. Monahan[4,5,6], Christopher J. Burns[4,5], Chunaram Choudhary[3], Marnie E. Blewitt[4,5], Anne K. Voss[4,5], Tim Thomas[4,5] & Mark A. Dawson[1,2,15,16]*

Acute myeloid leukaemia (AML) is a heterogeneous disease characterized by transcriptional dysregulation that results in a block in differentiation and increased malignant self-renewal. Various epigenetic therapies aimed at reversing these hallmarks of AML have progressed into clinical trials, but most show only modest efficacy owing to an inability to effectively eradicate leukaemia stem cells (LSCs)[1]. Here, to specifically identify novel dependencies in LSCs, we screened a bespoke library of small hairpin RNAs that target chromatin regulators in a unique ex vivo mouse model of LSCs. We identify the MYST acetyltransferase HBO1 (also known as KAT7 or MYST2) and several known members of the HBO1 protein complex as critical regulators of LSC maintenance. Using CRISPR domain screening and quantitative mass spectrometry, we identified the histone acetyltransferase domain of HBO1 as being essential in the acetylation of histone H3 at K14. H3 acetylated at K14 (H3K14ac) facilitates the processivity of RNA polymerase II to maintain the high expression of key genes (including *Hoxa9* and *Hoxa10*) that help to sustain the functional properties of LSCs. To leverage this dependency therapeutically, we developed a highly potent small-molecule inhibitor of HBO1 and demonstrate its mode of activity as a competitive analogue of acetyl-CoA. Inhibition of HBO1 phenocopied our genetic data and showed efficacy in a broad range of human cell lines and primary AML cells from patients. These biological, structural and chemical insights into a therapeutic target in AML will enable the clinical translation of these findings.

AML is organized in a loose hierarchy, in which a small population of self-renewing LSCs give rise to a large population of more-mature leukaemic blasts[2]. Although several human and mouse AML cell lines have undergone chemical and genetic screens to identify targetable dependencies in this disease[3–5], the majority of these models do not replicate the functional properties of LSCs. Analogous to the effective maintenance of embryonic stem cells with therapeutic pressure to decrease differentiation[6], a method to sustain cells with the transcriptional and functional properties of LSCs in liquid culture has previously been established[7]. Importantly, we concurrently established an isogenic population of AML blasts.

Because regulators of transcription are the most frequent mutational targets in AML[8], we performed a pooled negative-selection screen with a customized small hairpin RNA (shRNA) library against 270 known chromatin modifiers to identify transcriptional regulators that are required for the maintenance of functionally validated LSCs[7].

[1]Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. [2]Sir Peter MacCallum Department of Oncology, University of Melbourne, Melbourne, Victoria, Australia. [3]Department of Proteomics, The Novo Nordisk Foundation Center for Protein Research, Faculty of Health and Medical Sciences, University of Copenhagen, Copenhagen, Denmark. [4]The Walter and Eliza Hall Institute of Medical Research, Melbourne, Victoria, Australia. [5]The Department of Medical Biology, The University of Melbourne, Melbourne, Victoria, Australia. [6]Cancer Therapeutics CRC, Melbourne, Victoria, Australia. [7]Medicinal Chemistry Theme, Monash Institute of Pharmaceutical Sciences, Monash University, Melbourne, Victoria, Australia. [8]Commonwealth Scientific and Industrial Research Organisation (CSIRO), Biomedical Program, Parkville, Victoria, Australia. [9]Discovery Biology, Griffith Institute for Drug Discovery, Griffith University, Nathan, Queensland, Australia. [10]Children's Cancer Institute, Kensington, New South Wales, Australia. [11]School of Pharmaceutical Sciences, Nanjing Tech University, Nanjing, China. [12]Wellcome-MRC Cambridge Stem Cell Institute, Department of Haematology, University of Cambridge, Cambridge, UK. [13]Haematological Cancer Genetics, Wellcome Sanger Institute, Cambridge, UK. [14]Institute for Medical Biochemistry, University of Veterinary Medicine Vienna, Vienna, Austria. [15]Centre for Cancer Research, University of Melbourne, Melbourne, Victoria, Australia. [16]Department of Haematology, Peter MacCallum Cancer Centre, Melbourne, Victoria, Australia. *e-mail: Mark.Dawson@petermac.org

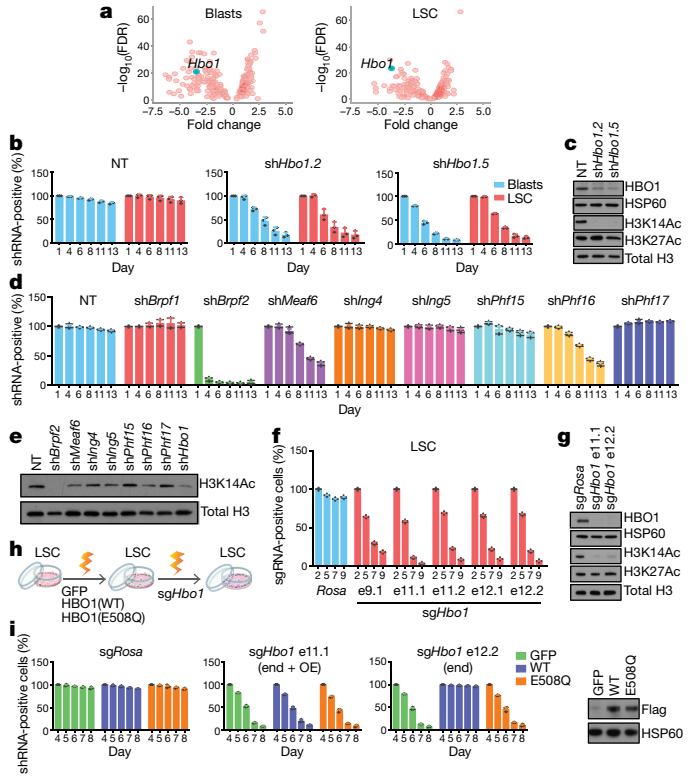


Fig. 1 | HBO1 is an essential dependency in LSCs. a, Pooled negative-selection screening in MLL-AF9 bulk blasts (left) and LSCs (right). Volcano plot depicts the changes in representation of shRNAs in the screen between day 2 and day 14. Each dot represents the mean of two independent experiments for a gene. **b**, Negative-selection competition assays, depicting the percentage over time of blasts and LSCs that express *Hbo1* shRNAs. $n = 3$ experiments. Mean + s.e.m. NT, non-targeting shRNA. **c**–**e**, Immunoblot of HBO1, H3K14ac and H3K27ac ($n = 3$) (**c**), negative-selection competition assays ($n = 3$; mean + s.e.m.) (**d**) and immunoblot of H3K14ac in LSCs that express shRNAs that target members of the Hbo1 complex ($n = 3$) (**e**). **f**, **g**, Negative-selection competition assays ($n = 3$; mean + s.e.m.) (**f**) and immunoblot of HBO1, H3K14ac and H3K27ac (representative of $n = 3$ biological replicates) (**g**) in LSCs that express Cas9 and sgRNAs against *Rosa26* (control) or the *Hbo1* catalytic domain. **h**, Schematic outline of HBO1 rescue experiments. Wild-type (WT) *Hbo1*, catalytic-mutant *Hbo1*^{E508Q} (resistant to sgRNA against *Hbo1* e12.2) or GFP were overexpressed in LSCs that express Cas9, and then transduced with sgRNAs targeting *Hbo1*. **i**, HBO1-rescue negative-selection competition assays ($n = 3$; mean + s.e.m.). sgRNA against *Hbo1* e11.1 targets endogenous (end) and overexpressed (OE) *Hbo1*. Right, immunoblot of overexpressed wild-type and catalytic-mutant HBO1.

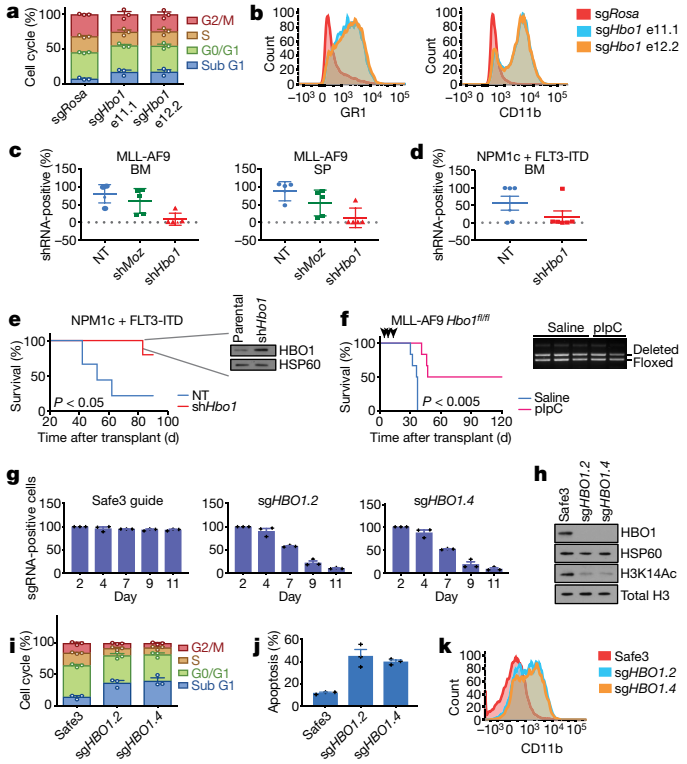


Fig. 2 | Loss of HBO1 impairs LSC maintenance in vivo. a, **b**, Cell-cycle profile ($n = 3$; mean + s.e.m.) (**a**) and surface expression of GR1 and CD11b ($n = 3$) (**b**) in LSCs that express Cas9, transduced with *Hbo1* sgRNAs. **c**, Percentage of shRNA-positive cells in bone marrow (BM) and spleen (SP) at end point, in recipients transplanted with quinary MLL-AF9 cells. Mean ± s.d. $n = 5$ mice per group. **d**, Percentage of shRNA-positive cells (mean ± s.e.m.) in bone marrow at end point, in recipients transplanted with tertiary NPM1c and FLT3-ITD cells. $n = 6$ mice per group. **e**, Kaplan–Meier curves of recipients injected with NPM1c and FLT3-ITD cells that express shRNA targeting *Hbo1*. $n = 5$ mice per group. Inset, immunoblot of HBO1 in NPM1c and FLT3-ITD cells expressing *Hbo1* shRNA from a recipient that died showed that the hairpin was inactivated. *Hbo1* levels from parental NPM1c and FLT3-ITD cells are shown for comparison. **f**, Kaplan–Meier curves of C57BL/6 mice injected with 10^3 *Hbo1*$^{flox/flox}$ *Mx1-cre* primary MLL-AF9 leukaemic cells. Mice were treated with saline or pIpC (arrowheads). Inset, representative genotyping of recipient bone marrow at end point. $n = 6$ mice per group. **g**–**k**, Negative-selection competition assays ($n = 3$; mean + s.e.m.) (**g**), immunoblot of HBO1 and H3K14ac ($n = 3$) (**h**), cell-cycle profile ($n = 3$; mean + s.e.m.) (**i**), apoptosis ($n = 3$; mean + s.e.m.) (**j**) and surface expression of CD11b ($n = 3$) (**k**) in Molm13 cells that express Cas9 and *HBO1* sgRNAs. Safe3 guide was used as a negative control.

The screen was highly reproducible, and clearly identified shared and unique dependencies in LSCs and AML blasts (Extended Data Fig. 1a). We observed far fewer dependencies in LSCs: less than one-third of the shRNAs were depleted in the LSCs compared to the blasts (Fig. 1a, Supplementary Table 1). It has previously been shown that the BET bromodomain proteins (BRD2, BRD3 and BRD4) were not a major dependency in this LSC model[7], and—in addition—we found that most of the hitherto-identified epigenetic dependencies (including DOT1L, LSD1, EZH2 and PRMT5) that have been the focus of clinical therapies[1] selectively eradicate only the blasts and not the LSCs (Extended Data Fig. 1b, c). Of the few dependencies identified in the LSCs, we chose to focus on *Hbo1* (also known as *Kat7*) as it is not a recognized essential gene and was equally effective in eradicating the blast and LSC populations (Fig. 1a, b, Extended Data Fig. 1d). HBO1 is one of five mammalian members of the highly conserved MYST family of acetyltransferases. Recent efforts to identify unique and global genetic dependencies in human cells have highlighted the fact that *MOF* (also known as *KAT8*) and *TIP60* (also known as *KAT5*) are pan-essential genes[9], whereas *HBO1* is highly expressed in human AML (Extended Data Fig. 2), in which it shows a clear and unique dependency[9] (Fig. 1b, c, Extended Data Fig. 3a, b). HBO1 has previously been reported to function as a major transcriptional regulator, primarily via histone acetylation, and—although various histone modifications have been attributed to HBO1[10–12]—these conflicting reports are probably influenced by the specificity of the antibodies used. Therefore, to precisely identify the major histone modifications regulated by HBO1, we coupled conditional deletion of HBO1 in AML cells with quantitative mass spectrometry[13]. These data clearly demonstrate that the acetylation of histone H3 at K14 is the major non-redundant chromatin modification that is mediated by HBO1 (Fig. 1c, Supplementary Table 2).

Similar to most histone acetyltransferases, HBO1 can interact with several scaffolding proteins to form functionally distinct catalytically active complexes[14]. Therefore, to identify the major complex members required for the maintenance of LSCs, we assessed the functional effect

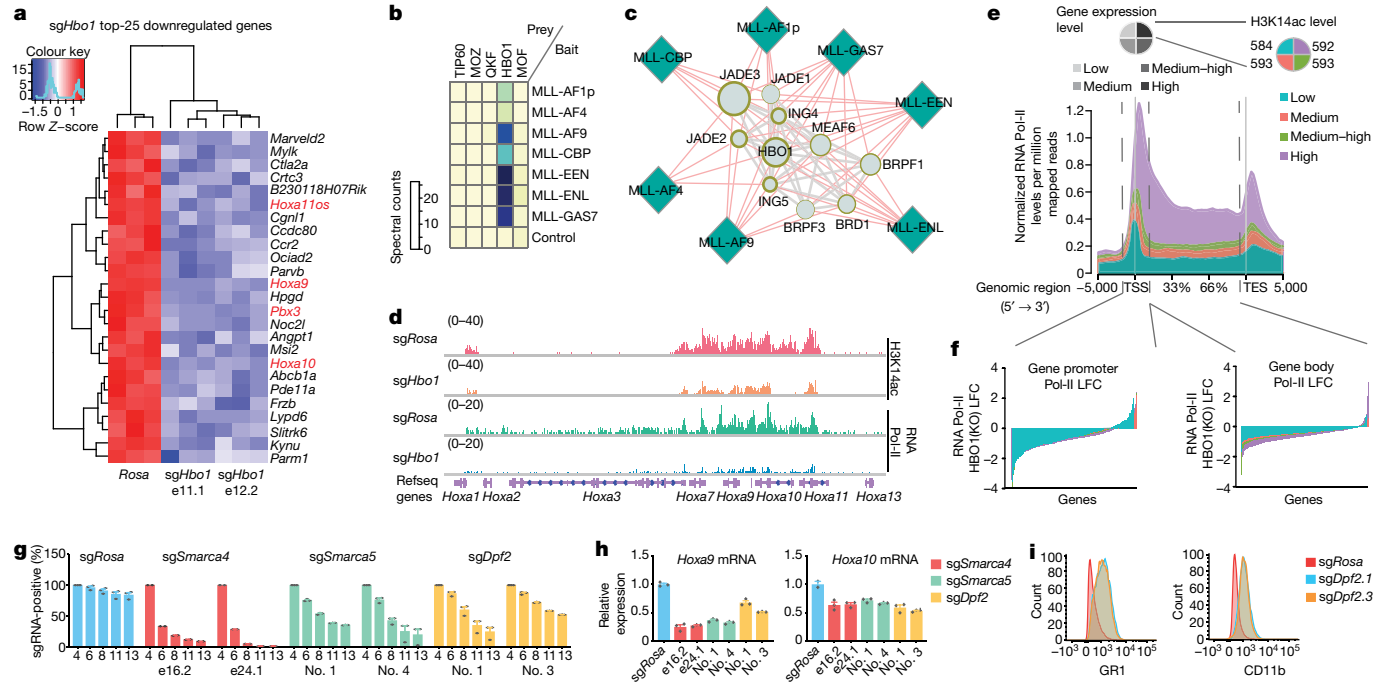


Fig. 3 | HBO1 regulates expression of the 5′ end of the HOXA cluster in AML. **a**, Heat map depicting normalized expression of the top-25 downregulated genes in LSCs with HBO1 loss. **b**, Heat map displaying the mean spectral counts from affinity-purification mass spectrometry data of seven MLL1 translocation partners for all members of the MYST family. **c**, Network analysis of known protein interaction partners of HBO1 (thick grey lines), linking to known members of the HBO1 complex and HBO1 network to seven MLL fusions (thin red line). Node size represents the mean spectral counts for each interaction. The width of the node border is scaled to the number of interactions with the MLL fusions. **d**, ChIP–seq profiles of H3K14ac and RNA Pol-II at the HOXA gene cluster in LSCs that express Cas9 and *Hbo1* sgRNA. Representative of $n = 3$ biological replicates. **e**, RNA Pol-II coverage across highly expressed genes (high) divided according to H3K14ac levels. TSS, transcription start site. **f**, Waterfall plots of change in RNA Pol-II (fold change, expressed in $\log_2$, (LFC) of RNA Pol-II) binding throughout the gene promoter (−30 to +300 bp) and gene body (+300 bp to transcription end site (TES)) following HBO1 loss (knockout (KO)) for highly expressed genes, divided according to H3K14ac levels. **g**, **h**, Negative-selection competition assays ($n = 3$; mean + s.e.m.) and *Hoxa9* and *Hoxa10* mRNA expression ($n = 3$; mean + s.e.m.) (**h**) in LSCs that express Cas9 and *Smarca4*, *Dpf2* or *Smarca5* sgRNAs. **i**, Surface expression of GR1 and CD11b in LSCs that express Cas9 and *Dpf2* sgRNAs. $n = 2$.

of genetic depletion of the members of the HBO1 complex. We found that knockdown of BRPF2 (also known as BRD1), MEAF6 and PHF16 (also known as JADE3) phenocopied the functional and biochemical effects of HBO1 reduction (Fig. 1d, e, Extended Data Fig. 3c). As all of the components of the HBO1 complex that resulted in impaired maintenance of LSCs also caused a global reduction in H3K14ac, we reasoned that the catalytic domain of HBO1 might be the critical target. CRISPR domain screening with five separate single-guide RNAs (sgRNAs) against the MYST domain confirmed that this domain was essential for H3K14ac and LSC survival (Fig. 1f, g). Although CRISPR domain screening is an effective approach to identifying functional domains for drug discovery[15], our rescue experiments with wild-type and catalytically inactive HBO1 provided the highest level of confidence that the histone acetyltransferase domain of HBO1 was the critical therapeutic target in the complex (Fig. 1h, i).

The dominant cellular phenotypes that resulted from the loss of HBO1 included an induction of apoptosis, a prominent G0/G1 cell-cycle arrest and a marked differentiation of the immature LSC population (Fig. 2a, b, Extended Data Fig. 3d–h). These data highlighted the importance of HBO1 in LSC maintenance in an ex vivo model system. To address the broader application of our findings in the absence of therapeutic pressure to maintain the LSC state[7], we generated an enriched population of LSCs in an in vivo mouse model[16] and performed a competition assay to assess the requirement of HBO1 for LSC maintenance in vivo. Here, we transplanted a fixed ratio of 90% shRNA-expressing cells and followed the percentage of shRNA-expressing cells that contributed to the leukaemia in vivo. Despite 90% of LSCs expressing *Hbo1* shRNA being transplanted, less than 5% of them remain at the time of death from leukaemia—demonstrating a marked negative selection (Fig. 2c). By contrast, both the non-targeting shRNA and shRNA against the closely related member of the MYST family *Moz* (also known as *Kat6a*) show no detrimental effect to LSCs. Similar results were also seen in the NPM1c and FLT3-ITD mouse model[17] (Fig. 2d). Moreover, the mice transplanted with shRNAs against *Hbo1* showed a substantial survival benefit (Fig. 2e, Extended Data Fig. 4a), raising the prospect that *Hbo1*-null LSCs are incapable of perpetuating the disease.

To explore this possibility further, we generated leukaemias using an MX1–Cre model for the conditional deletion of *Hbo1* (Extended Data Fig. 4b). The resulting leukaemia was then transplanted into secondary recipient mice and polyinosinic:-polycytidylic acid (pIpC) was administered after engraftment. Neither pIpC injection nor heterozygous deletion of *Hbo1* affect survival or leukaemia latency (Extended Data Fig. 4c). Leukaemic cells derived from $Hbo1^{flox/flox}$ *Mx1-cre* mice show a marked survival advantage, and none of the fatal leukaemia that occurred in pIpC-treated mice showed complete loss of HBO1 (Fig. 2f). By contrast, homozygous deletion of *Moz* showed no effects on survival in two separate mouse models of AML (Extended Data Fig. 4d–f). Together, these data confirm the results from our ex vivo model, and provide compelling evidence that HBO1 is an essential requirement for LSC maintenance.

To assess the generality of our findings beyond mouse models of AML, we chose to delete HBO1 using CRISPR–Cas9 in a range of human AML cell lines that encompasses a variety of oncogenic drivers prevalent in AML[8]. We found the majority of AML cell lines recapitulate our results in the mouse LSCs, and show an impaired survival of *HBO1*-deleted cells that results from an induction of apoptosis, a G0/G1 cell-cycle arrest

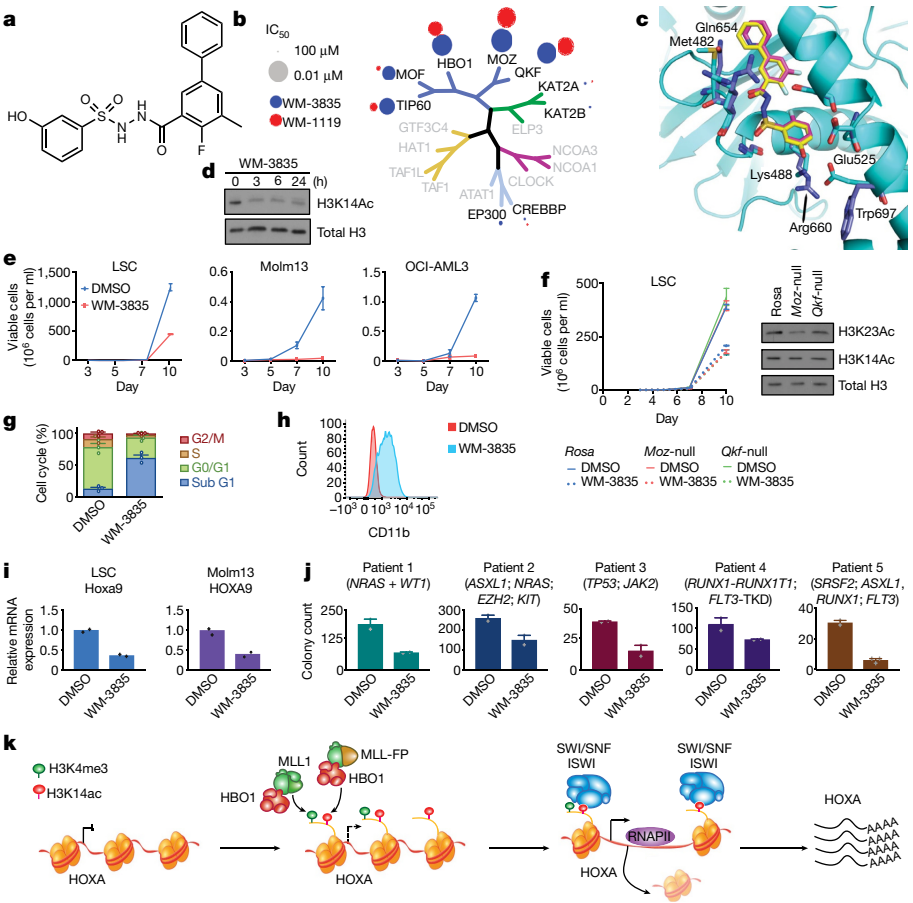


Fig. 4 | Treatment with WM-3835 reduces AML growth. a, Chemical structure of WM-3835. **b**, Selectivity and potency profile of WM-3835 and WM-1119, superimposed on the KAT family dendrogram, as measured by a histone acetyltransferase inhibition assay. Values are listed in Extended Data Fig. 8a. **c**, Ribbon representation of the HBO1–BPRF2 crystal structure (cyan) with WM-3835 bound (yellow with element colouring; RCSB Protein Data Bank code (PDB) 6MAJ) overlaid with the corresponding region of MYST^{CRYST} with WM-1119 bound (magenta with element colouring; PDB 6BA4), in which the non-conserved residues of MYST^{CRYST} are shown in (blue). The key residues highlighted show a hydrogen bond with the phenol of WM-3835. **d**, Immunoblot of H3K14ac in LSCs treated with WM-3835. $n = 2$. **e**, LSCs and human AML cell proliferation with 1 μM WM-3835. $n = 3$. Mean + s.e.m. **f**. Proliferation of LSCs with *Moz* or *Qkf* (also known as *Kat6b* and *Morf*) deletion with WM-3835. $n = 3$. Mean ± s.e.m. Right, immunoblot of H3K23ac and H3K14ac in LSCs with *Moz* or *Qkf* deletion. $n = 2$. **g**, **h**, Cell-cycle profile ($n = 3$; mean + s.e.m.) (**g**) and surface expression of CD11b ($n = 3$) (**h**) in Molm13 cells with WM-3835. **i**. *Hoxa9* expression in mouse LSCs and *HOXA9* expression in Molm13 cells treated with WM-3835. $n = 2$. **j**, Clonogenic assays with primary AML cells from patients, treated with WM-3835. $n = 5$. Mean + s.e.m. Gene symbols in parentheses denote mutated myeloid genes. **k**, Schematic overview of AML gene regulation by the HBO1 complex.

and prominent differentiation (Fig. 2g–k, Extended Data Fig. 5). By contrast, very few non-AML cell lines show a similar dependency on HBO1 (Extended Data Fig. 6). Having established the requirement of HBO1 in mouse and human models of AML, we next wanted to understand the molecular events that underpin the role of HBO1 in LSC maintenance. Consistent with the major cellular phenotype of myeloid differentiation, we found that HBO1 loss results in the marked enrichment of a myeloid differentiation gene-expression program (Extended Data Fig. 7a). The established role of HBO1 as a facilitator of transcription led us to examine the top downregulated genes after HBO1 deletion. These downregulated genes are some of the most highly expressed (Extended Data Fig. 7b) and include several homeobox genes (Fig. 3a), which are known to be important in LSC maintenance and are commonly upregulated in AML with a poor prognosis[18]. The requirement of HBO1 to sustain the expression of the essential LSC genes within the 5′-HOXA gene cluster is conserved in human AML cells (Extended Data Fig. 7c), and the dominant role of these genes in mediating the cellular phenotypes of HBO1 loss is highlighted by the fact that overexpression of *Hoxa9* or *Hoxa10* considerably rescues the myeloid differentiation and loss of viability that is observed after depletion of HBO1 (Extended Data Fig. 7d-f).

Many of the genes that are downregulated after loss of HBO1—particularly the homeobox genes—are established targets of both wild-type MLL1 and MLL1 fusion proteins[19]. Using quantitative proteomics in an isogenic leukaemia cell line designed to express a single copy of seven distinct MLL1 fusion proteins[20], we identified members of the HBO1 complex that are functionally required to maintain LSCs (Fig. 1d) as a strong interactors with the N terminus of MLL1 (Fig. 3b, c). Although these findings provide molecular insights into how the HBO1 complex is recruited to specific gene loci, to further understand the role of HBO1 in regulating these genes we performed chromatin immunoprecipitation with sequencing (ChIP–seq) analyses for H3K14ac and RNA polymerase II (RNA Pol-II). These data show that H3K14ac deposited by HBO1 is widespread throughout the genome, but that at the highly expressed genes repressed by loss of HBO1, H3K14ac and RNA Pol-II blanket the entire coding region of the gene (Fig. 3d, Extended Data Fig. 7g). H3K14ac is an evolutionarily conserved histone modification, and recent evidence suggests that H3K14ac may regulate transcriptional elongation[21]. Consistent with this, we find markedly increased levels of RNA Pol-II within the coding region of highly expressed genes that contain the highest levels of H3K14ac (Fig. 3e). Furthermore, expressed genes with the highest level of H3K14ac have

the lowest RNA Pol-II travelling ratio, and loss of HBO1 leads to a more prominent loss of RNA Pol-II within the body of these genes (Fig. 3f, Extended Data Fig. 7h). The processivity of RNA Pol-II is greatly facilitated by chromatin remodelling complexes, and H3K14ac has previously been shown to be specifically bound by SMARCA4[22], DPF2[23] and members of the ISWI family[24], resulting in the marked potentiation of their remodelling activity. The members of these chromatin remodelling complexes show a similar cancer-cell-line dependency profile to HBO1 (with a predilection for AML[9]) and also phenocopy the effects of HBO1 loss in LSCs (Fig. 3g–i, Extended Data Fig. 7i).

Our genetic data in both mouse and human AML cells clearly identified the catalytic activity of HBO1 as the central therapeutic target. A long-standing challenge in the field has been to develop highly selective small-molecule histone-acetyltransferase inhibitors that discriminate between the major families of histone acetyltransferases. It has recently been demonstrated that the acylsulfonylhydrazide backbone provides a simple chemical scaffold for the generation of selective inhibitors of MYST-family acetyltransferases[25]. Using this template, we generated WM-3835 (*N*′-(4-fluoro-5-methyl-[1,1′-biphenyl]-3-carbonyl)-3-hydroxybenzenesulfonohydrazide) (Fig. 4a), which retains specificity for the MYST acetyltransferases but has increased potency against HBO1 compared to WM-1119 (Fig. 4b, Extended Data Fig. 8a). We solved the crystal structure of HBO1 with WM-3835 bound in the acetyl-CoA binding site at 2.14 Å (Fig. 4c, Extended Data Fig. 9). Overlaying this crystal structure with WM-1119 in complex with a modified MYST acetyltransferase domain[25] ($MYST^{CRYST}$) shows that WM-3835 makes additional interactions with the protein surface, which may explain the increased activity of WM-3835 against HBO1. Specifically, the WM-3835 phenol forms a hydrogen-bonding network with Glu525 and Lys488, neither of which is conserved throughout the MYST family.

WM-3835 is a cell-permeable small molecule that results in a rapid and selective reduction in levels of H3K14ac (Fig. 4d). Treatment of a diverse set of AML cell lines with WM-3835 resulted in a marked reduction in tumour-cell viability (Fig. 4e, Extended Data Fig. 10a) that was not observed after treatment with the inactive analogue WM-2474[25] (Extended Data Fig. 8b). Notably, we observed an excellent dose–response relationship between a reduction of levels of H3K14ac and cell viability (Extended Data Fig. 8c, d). Although WM-3835 retains potency against MOZ and QKF (also known as MORF or KAT6B), CRISPR–CAS9-mediated deletion of these enzymes does not alter the activity of WM-3835 (Fig. 4f), which highlights the fact that the efficacy of WM-3835 in AML is primarily via HBO1 inhibition. Moreover, treatment of cells with WM-3835 phenocopied the molecular and cellular effects of genetic depletion of HBO1 by inducing apoptosis, a G0/G1 cell-cycle arrest, differentiation of human AML cells and transcriptional repression on *HOXA9* and *HOXA10* (Fig. 4g, i, Extended Data Fig. 8e–h). Similar to our genetic studies, overexpression of *Hoxa9* and *Hoxa10* ameliorated the effects of WM-3835 (Extended Data Fig. 8i). Although the rapid metabolism—including glucuronidation of WM-3835—precluded efficacy experiments in vivo (Extended Data Fig. 10b-c), the compound showed a prominent reduction of clonogenic potential in primary human AML cells (derived from several patients) that contained different driver mutations, highlighting the therapeutic potential of catalytic inhibitors against HBO1 in AML (Fig. 4j).

Central to the ambition to alter the natural history of AML is the requirement for new therapies that effectively target LSCs from the outset. LSCs serve as the reservoir for evolving resistance to conventional and targeted therapies, and our clinical experience has clearly proven that monotherapies are incapable to subvert the vast adaptive potential of LSCs. Therefore, the future lies in identifying key therapeutic targets in LSCs that can be leveraged in combination with other effective agents, including conventional chemotherapy. Here we identify HBO1 as a targetable dependency in LSCs. Our molecular insights suggest that MLL1 recruits HBO1 to regulate highly expressed LSC genes (including the HOXA cluster) through H3K14ac, which potentiates the activity of specific chromatin remodelling complexes and enables a greater processivity of RNA Pol-II (Fig. 4k). The blueprint for selective and potent inhibition of HBO1, together with these biological insights, provide the impetus and platform for the translation of these findings into the clinical setting.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1835-6.

1. Dawson, M. A. The cancer epigenome: concepts, challenges, and therapeutic opportunities. *Science* **355**, 1147–1152 (2017).
2. Thomas, D. & Majeti, R. Biology and relevance of human acute myeloid leukemia stem cells. *Blood* **129**, 1577–1585 (2017).
3. Wang, T. et al. Gene essentiality profiling reveals gene networks and synthetic lethal interactions with oncogenic Ras. *Cell* **168**, 890–903 (2017).
4. Tzelepis, K. et al. A CRISPR dropout screen identifies genetic vulnerabilities and therapeutic targets in acute myeloid leukemia. *Cell Reports* **17**, 1193–1205 (2016).
5. Zuber, J. et al. RNAi screen identifies Brd4 as a therapeutic target in acute myeloid leukaemia. *Nature* **478**, 524–528 (2011).
6. Ying, Q. L. et al. The ground state of embryonic stem cell self-renewal. *Nature* **453**, 519–523 (2008).
7. Fong, C. Y. et al. BET inhibitor resistance emerges from leukaemia stem cells. *Nature* **525**, 538–542 (2015).
8. Grimwade, D., Ivey, A. & Huntly, B. J. Molecular landscape of acute myeloid leukemia in younger adults and its clinical relevance. *Blood* **127**, 29–41 (2016).
9. Tsherniak, A. et al. Defining a cancer dependency map. *Cell* **170**, 564–576 (2017).
10. Kueh, A. J., Dixon, M. P., Voss, A. K. & Thomas, T. HBO1 is required for H3K14 acetylation and normal transcriptional activity during embryonic development. *Mol. Cell. Biol.* **31**, 845–860 (2011).
11. Doyon, Y. et al. ING tumor suppressor proteins are critical regulators of chromatin acetylation required for genome expression and perpetuation. *Mol. Cell* **21**, 51–64 (2006).
12. Sauer, T. et al. MYST2 acetyltransferase expression and histone H4 Lysine acetylation are suppressed in AML. *Exp. Hematol.* **43**, 794–802 (2015).
13. Weinert, B.T. et al. Time-resolved analysis reveals rapid dynamics and broad scope of the CBP/p300 acetylome. *Cell* **174**, 231–244 (2018).
14. Saksouk, N. et al. HBO1 HAT complexes target chromatin throughout gene coding regions via multiple PHD finger interactions with histone H3 tail. *Mol. Cell* **33**, 257–265 (2009).
15. Shi, J. et al. Discovery of cancer drug targets by CRISPR–Cas9 screening of protein domains. *Nat. Biotechnol.* **33**, 661–667 (2015).
16. Krivtsov, A. V. et al. Transformation from committed progenitor to leukaemia stem cell initiated by MLL-AF9. *Nature* **442**, 818–822 (2006).
17. Mupo, A. et al. A powerful molecular synergy between mutant nucleophosmin and Flt3-ITD drives acute myeloid leukemia in mice. *Leukemia* **27**, 1917–1920 (2013).
18. Alharbi, R. A., Pettengell, R., Pandha, H. S. & Morgan, R. The role of HOX genes in normal hematopoiesis and acute leukemia. *Leukemia* **27**, 1000–1008 (2013).
19. Ballabio, E. & Milne, T. A. Epigenetic control of gene expression in leukemogenesis: cooperation between wild type MLL and MLL fusion proteins. *Mol. Cell. Oncol.* **1**, e955330 (2014).
20. Skucha, A. et al. MLL-fusion-driven leukemia requires SETD2 to safeguard genomic integrity. *Nat. Commun.* **9**, 1983 (2018).
21. Church, M., Smith, K. C., Alhussain, M. M., Pennings, S. & Fleming, A. B. Sas3 and Ada2(Gcn5)-dependent histone H3 acetylation is required for transcription elongation at the de-repressed *FLO1* gene. *Nucleic Acids Res.* **45**, 4413–4430 (2017).
22. Shen, W. et al. Solution structure of human Brg1 bromodomain and its specific binding to acetylated histone tails. *Biochemistry* **46**, 2100–2110 (2007).
23. Huber, F. M. et al. Histone-binding of DPF2 mediates its repressive role in myeloid differentiation. *Proc. Natl Acad. Sci. USA* **114**, 6016–6021 (2017).
24. Dann, G. P. et al. ISWI chromatin remodellers sense nucleosome modifications to determine substrate preference. *Nature* **548**, 607–611 (2017).
25. Baell, J. B. et al. Inhibitors of histone acetyltransferases KAT6A/B induce senescence and arrest tumour growth. *Nature* **560**, 253–257 (2018).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized and investigators were not blinded to allocation during experiments and outcome assessment.

Cell culture

MLL–AF9 bulk blasts and LSCs were generated as previously described[7]. Mouse and human cell lines (NOMO-1, Molm13, MV4;11, HL-60, OCI-M2, OCI-AML3, K562, NB4, SKM-1 and KG-1) were maintained in RPMI-1640 supplemented with 10% FCS, 2 mM GlutaMAX, 100 IU ml^{-1} penicillin, 100 μg ml^{-1} streptomycin under standard culture conditions (5% CO_2, 37 °C). Blasts and LSCs were maintained in the presence of 0.1% DMSO or 1 μM I-BET151, respectively, and IL-3 (10 ng ml^{-1}). HEK293T cells were maintained in DMEM supplemented with 10% FCS, 100 IU ml^{-1} penicillin, 100 μg ml^{-1} streptomycin in 10% CO_2 at 37 °C. All cell lines were regularly tested and verified to be mycoplasma-negative by PCR analysis by in-house genotyping. Human cell lines were authenticated by short tandem repeat profiling through the Australian Genome Research Facility.

Virus production and transduction

Retrovirus was produced by triple transfection of HEK293T cells with a retroviral LMP-blue fluorescent protein (BFP) transfer vector and structural pMD1-gag-pol plasmid and Vsv-g envelope plasmid at a 0.75:0.22:0.03 ratio, as previously described[26]. Lentivirus was produced by triple transfection of HEK293T cells with a lentiviral transfer vector, and the packaging plasmids psPAX2 and Vsv-g at a 0.5:0.35:15 ratio. All transfections were performed using polyethylenimine (PEI). Viral supernatants were collected 48 h after transfection, filtered through a 0.45-μm filter and added to target cells.

To examine the effect of loss of MOZ function on progression of leukaemia, fetal liver cells were isolated from embryos with a germline deletion of *Moz* or littermate controls[27] at embryonic day (E)13. E13 embryos were used because *Moz*-null embryos die by E14. Fetal liver cells (C57B/6; CD45.2 cell-surface phenotype) were transfected with MSCV expressing either MLL-AFP and GFP or MEIS1, HOXA9 and GFP or control viruses: empty vector (GFP) or MEIS1–GFP or HOXA9–GFP alone (not shown); and prepared for infection as in the previous paragraph, except that the ECO envelope protein was used. After overnight culture, infected fetal liver cells were injected into C57B/6 CD45.1 recipient mice, which had been irradiated with a single dose of 700 rads.

Pooled negative-selection RNA interference screening

A custom shRNA library targeting 270 mouse epigenetic enzymatic genes was designed using the Designer of Small Interfering RNA website (http://biodev.cea.fr/DSIR/DSIR.html) and subcloned into the LMP-BFP vector with selectable markers eBFP or puromycin, as previously described[28]. After sequence verification, 1,922 shRNAs (6–8 per gene) were combined with several positive- and negative-control shRNAs at equal concentration in one pool. This pool was used to produce retrovirus, which was then transduced into 4×10^6 MLL–AF9 bulk blasts and LSCs at a multiplicity of infection of 0.3 and selected with 3 and 5 μg ml^{-1} puromycin, respectively, commencing 30 h after transduction. Throughout 14 days of puromycin selection, more than 20 million cells were maintained at each passage to preserve 10,000-fold library representation. Genomic DNA from day 2 to day 14 was isolated (DNeasy Blood & Tissue Kit, Qiagen) from both blasts and LSCs. shRNA sequences were amplified by PCR with primers containing adaptors for Illumina sequencing, as previously described[29]. The resulting libraries were sequenced with single-end 50-bp reads on a HiSeq2500. The shRNA sequences were mapped to the shRNAs within the pool, and the shRNA counts were analysed as previously described[30]. The likelihood ratio test was used to determine the hairpins that were significantly depleted over the time course of the experiment. Genes with at least two hairpins depleted by greater than tenfold were considered to be significant dependencies.

CRISPR–Cas9-mediated gene disruption

sgRNA oligonucleotides (Sigma-Aldrich) were phosphorylated, annealed and cloned into lentiviral expression vectors, pKLV-U6gRNA(BbsI)-PGKpuro2ABFP (Addgene 50946, deposited by K. Yusa). Cells were first transduced with the FUCas9Cherry (Addgene 70182, deposited by M. Herold) and sorted by fluorescence-activated cell sorting for high mCherry expression, and then subsequently transduced with the pKLV sgRNA expression vector.

shRNA and sgRNA competitive proliferation assay

Bulk blasts and LSCs were transduced with retrovirus expressing a gene-specific shRNA, the percentage of BFP-positive cells was measured between day 1 and day 13 after transduction, and normalized to the percentage of BFP positive cells at day 1. For sgRNA, Cas9-expressing cells were transduced with a lentivirus expressing an sgRNA targeting *Hbo1* and the percentage of double-positive BFP and mCherry cells was measured between day 2 and day 14 after transduction, and normalized to the percentage of $BFP^+mCherry^+$ cells at day 2 or day 4. All shRNA and sgRNA sequences are provided in Supplementary Table 3.

Antibodies

Antibodies used for immunoblotting were rabbit anti-acetyl-histone H3 (Lys14) (D4B9, Cell Signalling Technology), mouse anti-histone H3K14ac (13HH3-1A5, Active Motif), rabbit anti-histone H3 (acetyl K27) (ab4729, Abcam), rabbit anti-KAT7/HBO1/MYST2 (ab70183, Abcam), rabbit anti-histone H3 (ab1791, Abcam), mouse anti-HSP60 (C10, Santa Cruz), rabbit anti-HSP60 (H-300, Santa Cruz), mouse anti-Flag (M2, Sigma), mouse anti-RNA polymerase II (CTD4H8, Millipore). Antibodies used for flow cytometry were Alexa Fluor 700 anti-GR1 (108422, BioLegend) and Brilliant Violet 605 anti-CD11b (101237, Biolegend) and APC/Cy7 anti-mouse CD117 (also known as KIT) (313228, Biolegend).

Flow cytometry

For cell apoptosis, shRNA- or sgRNA-positive cells were washed once with PBS and assessed using FITC-conjugated annexin V (640906, Biolegend) and 4′,6-diamidino-2-phenylindole (DAPI) (D9542, Sigma) staining, according to the manufacturer's instructions. For cell-cycle analysis, shRNA- or sgRNA-positive cells were washed with PBS and fixed for at least 2 h at −20°C in 70% ethanol. Fixed cells were washed with PBS and incubated at 4 °C in DAPI staining solution (1 mg ml^{-1} DAPI, 0.05% (v/v) Triton X-100 in PBS) for 30 min. For surface expression of myeloid markers, shRNA- or sgRNA-positive cells were washed in PBS, and stained for GR1 or CD11b on ice for 30 min in PBS plus 5% FCS. All flow cytometry analyses were performed on a LSR Fortessa X-20 flow cytometer (BD Biosciences) and all data were analysed with FlowJo. Cell sorting was performed on a FACSAria Fusion 5 (BD Biosciences).

Immunoblotting

shRNA- or sgRNA-positive cells were lysed in 20 mM HEPES pH7.9, 0.5 mM EDTA, 2% SDS plus 1× protease inhibitor cocktail (Roche) by brief sonication. Lysates were heated to 95 °C in SDS sample buffer with 50 mM DTT for 5 min, separated by SDS–PAGE and transferred to PVDF membrane (Millipore). Membranes were blocked in 5% milk in TBS +0.1% Tween-20, probed with the indicated antibodies, and reactive bands were visualized using ECL Prime (GE).

Analysis of HBO1-regulated acetylation of core histones

Mouse MLL–AF9 cre^{ERT2} $Hbo1^{fl/fl}$ conditional knockout cells were labelled using stable isotope labelling by amino acids in cell culture, with 'light' ($^{12}C_6$,$^{14}N_4$-arginine and $^{12}C_6$,$^{14}N_2$-lysine) and 'heavy' (($^{13}C_6$,$^{15}N_4$-arginine and $^{13}C_6$,$^{15}N_2$-lysine, Cambridge Isotope Laboratories) amino acids. To delete HBO1, the heavy-labelled cells were treated with

4-hydroxytamoxifen (200 nM) for about 40 h, and light-labelled control cells were treated with vehicle control. Histones were extracted as previously described[31]. In brief, cells were lysed mechanically in ice-cold hypotonic lysis buffer (10 mM Tris pH 8.0, 1 mM KCl, 1.5 mM $MgCl_2$, 1 mM DTT and 1× complete protease inhibitor cocktail (Roche)) and intact nuclei were collected by centrifugation. Histones were acid-extracted with H_2SO_4 (0.4 N) and precipitated with TCA (33% final concentration). Purified histones from knockout and control cells were mixed in equal amounts, separated on SDS–PAGE and in-gel-digested with trypsin or LysC. (Sigma). The histone peptides were analysed by online nanoflow liquid-chromatography-coupled tandem mass spectrometry, using a Proxeon easy nLC system connected to a Q-Exactive HFX mass spectrometer (Thermo Scientific). The raw data were computationally processed using MaxQuant[32] (version 1.5.6.5) and searched against the UniProt database (downloaded 23 January 2014) using the integrated Andromeda search engine (http://www.maxquant.org/). The data were searched with three missed cleavages, a minimum peptide length of six amino acids and the requantify option selected, and acetylation was included as a variable modification.

Rescue assays

cDNA of *Hbo1* and *Hoxa10* were PCR-amplified from the cDNA library of mouse MLL–AF9 cells with primers containing a Flag. *Hoxa9* cDNA was amplified from pTRE rtTA Flag Hoxa9 GFP[33]. The catalytic mutant HBO1(E508Q) was generated by site-directed mutagenesis. Wild-type and mutant *Hbo1* were made resistant to *Hbo1* e12.2 sgRNA by silent point mutation of the protospacer-adjacent motif site corresponding to this sgRNA, using site-directed mutagenesis. All cDNAs were cloned into the lentiviral pHRSIN-P_{SFFV}-GFP-P_{PGK}-Puro vector[34]. LSCs expressing Cas9 were transduced with expression vectors and selected with 5 μg ml^{-1} puromycin for 1 week. Overexpression lines were then subsequently transduced with *Hbo1* sgRNA.

Mouse details

All mouse work was performed at the Peter MacCallum Cancer Centre animal facility, under approval E530 from the Peter MacCallum Cancer Centre animal ethics committee and at the Walter and Eliza Hall Institute of Medical Research with approval from the Walter and Eliza Hall Institute Animal Ethics Committee under approval 2015.015. MX1–Cre *Hbo1*$^{flox/flox}$ mice[10] and *Moz*$^{+/-}$ mice were as previously described[27].

In vivo competition assay

Quinary MLL–AF9 cells were transduced with non-targeting shRNA, *Hbo1* shRNA or *Moz* shRNA at 90% transduction efficiency. One hundred thousand cells were transplanted 48 h after transduction into 8-week-old female NOD/SCID/*Il2rg*-null (NSG) mice. BFP-positive shRNA-positive cells were determined by flow cytometry.

Leukaemia maintenance

The generation of *Mx1-cre Hbo1*$^{fl/fl}$ conditional knockout mice has previously been described[10]. KIT-positive cells from whole bone marrow were selected through magnetic bead selection (Miltenyi Biotec), and retrovirally transduced with the MSCV-MLL-AF9-IRES-YFP construct. Cells were transplanted in sublethally irradiated 6–8-week-old female C57BL/6 recipient mice. One hundred thousand leukaemic cells from the bone marrow were collected, and subsequently transplanted into sublethally irradiated 11-week-old female C57BL/6 recipient mice. Mice were randomized and pIpC (GE) was intraperitoneally administered 6, 10 and 14 days after transplantation, at 7.5 mg/kg. Amplification of wild-type and floxed alleles of leukaemic cells from bone marrow has previously been described[10].

RNA sequencing and analysis

RNA from sgRNA-positive cells was prepared using the Qiagen RNeasy kit. RNA concentration was quantified with a NanoDrop spectrophotometer (Thermo Scientific). Libraries were prepared using QuantaSeq 3′ mRNA Library Prep kit (Lexogen). Libraries were sequenced on a NextSeq500 with 75-bp single-end reads. All RNA sequencing experiments were performed in triplicate. Following trimming of poly-A tails with cutadapt[35] (v.1.14), reads were aligned to the mouse genome (ensembl_GRC38.78) using hisat2[36], and assigned to genes using htseq-count[37]. Differential gene expression analysis was performed using the edgeR[38] package in R (http://www.R-project.org/), and adjusted *P* values were calculated using the Benjamini–Hochberg method[39]. Genes with fold changes (expressed as log) below −1 and adjusted *P* values below 0.05 were considered to be significantly downregulated genes. Count data were voom-transformed using the voom function before performing gene-set testing with the mroast function[40], both from the limma package[41].

ChIP–seq and analysis

Ten to twenty million sgRNA-positive cells were crosslinked with 1% formaldehyde for 10 min at room temperature, and crosslinking was quenched by addition of 0.125 M glycine. Cells were then lysed in 1% SDS, 10 mM EDTA, 50 mM Tris-HCl pH 8.0 and protease inhibitors. Lysates were sonicated in a Covaris ultrasonicator to achieve a mean DNA fragment size of 500 bp. Immunoprecipitation with anti-H3K14ac (Cell Signalling Technolgies) or anti-RNA polymerase II (Millipore) was performed overnight at 4 °C in modified RIPA buffer (10 mM Tris-HCl pH 8.0, 90 mM NaCl, 1% Triton X-100 and 0.1% deoxycholate). Protein A or G magnetic beads (Life Technologies) were used to bind the antibody and associated chromatin. Reverse crosslinking of DNA was followed by DNA purification using the QIAquick PCR purification kit (Qiagen). Sequencing libraries were prepared from eluted DNA using ThruPLEX DNA-seq kit (Rubicon). Libraries were size-selected between 200–500 bp and sequenced on a NextSeq500 with 75-bp single-end reads. Following the removal of Illumina adaptors using cutadapt[35], reads were aligned to a joint reference genome of mouse (ensembl_GRCm38.78) and *Drosophila* (ensembl_BDGP5.78) using bwa-mem (v. 0.7.13). SAM files were converted to BAM files using samtools[42] (v. 1.4.1). A scaling factor was calculated using the *Drosophila* spike-in, as previously described[43]. The scaling factor was used to normalize the coverage across the genome, when calculated using bamCoverage from deepTools[44] (v. 2.5.3) with bin sizes of 10 bp, and filtered with ENCODE project ChIP blacklist regions for mm10 (https://www.encodeproject.org/annotations/ENCSR636HFF/). Genome-browser images were generated from the conversion of BAM files to TDF using igvtools[45] (v.2.3.95). Heat map plots were generated using deepTools[44] over the region 5 kb upstream to 5 kb downstream of the gene body of all genes. Coverage across the length of the gene body was scaled to 5 kb, and regions with no coverage were excluded from the plot.

Quantitative real-time PCR

RNA from sgRNA-positive cells 4–5 days after transduction, or cells treated with WM-3835 for 6–12 h, was extracted using the Qiagen RNAeasy kit. cDNA was prepared using SuperScript VILO (Life Technologies) according to the manufacturer's instructions. Quantitative real-time PCR was performed on an Applied Biosystems StepOnePlus using Fast SYBR green reagents (Thermo Scientific). Expression levels were determined using the $\Delta\Delta C_t$ method normalized to β2-microglobulin. All mRNA primer sequences are provided in Supplementary Table 3.

Cell proliferation assays

Cells were seeded at a constant density before treatment in triplicate, and treated with either 1 μM WM-3835, 1 μM WM-2474 or DMSO (0.1%) over the indicated time period. The drug was refreshed at least every two days. Cells were stained with DAPI and the live-cell number was calculated using the BD FACSVerse (BD Biosciences). To determine the half-maximal inhibitory concentration (IC_{50}) for the WM-3835, 4 h after seeding the cells at a constant density in duplicate, the cells were

treated with WM-3835, DMSO or positive control (3 μM puromycin) for 10 days. The drug and medium were refreshed at day 4 and day 7. At day 10, after incubating the cells with 600 μM of resazurin for 6 h, fluorescence was measured at λ_{ex} = 530 nm and λ_{em} = 590 nm, using a Microplate Reader (EnSpire, Perkin Elmer). Relative fluorescence units were converted to per cent of inhibition relative to controls on the same plate, and the data were fitted against a four-parameter logistic model to determine the IC_{50}.

Clonogenic assays in methylcellulose

Clonogenic potential was assessed through colony growth of bone marrow cells from patients with AML plated in cytokine-supplemented methylcellulose (MethoCult H4434, StemCell Technologies). Bone marrow was plated in duplicate at a cell dose of 2×10^4 cells per plate in the presence of vehicle (0.1% DMSO) or 1 μM WM-3835. Cells were incubated at 37 °C and 5% CO_2 for 12 days, at which time colonies were counted.

Patient material

Bone marrow containing >80% blasts was obtained from patients following consent and under full ethical approval by the Peter MacCallum Cancer Centre Research Ethics Committee (reference number: HREC/17/PMCC/69).

Lysine acetyltransferase biochemical assays

KAT enzymes were either produced or purchased, as previously described[25]. Lysine acetyltransferase assays were run as previously described[25], with two modification. First, 100 nM of full-length biotinylated histone H3 (for MOZ, QKF and HBO1) or histone H4 (for KAT5 and KAT8) proteins were used as the substrate, as indicated. Second, assays were run with 1 μM acetyl-CoA concentration, the approximate K_m for acetyl-CoA for these enzymes in this assay format.

HBO1 H3K14ac biomarker assay

The cell line U2OS was seeded at a density of 3,000 cells per well in 384-well optical-quality tissue-culture plates, in RPMI medium supplemented with 10% fetal bovine serum and 10 mM HEPES. The cells were allowed to adhere for 24 h under standard culture conditions (37 °C and 5% CO_2). At the end of this period, the cells were washed with medium. Compound dilutions prepared in DMSO were added to the medium, with negative-control wells reserved for treatment with DMSO-only and 100%-inhibition positive controls at 10 μM concentration. After incubation for 24 h, the cells were fixed with 4% formaldehyde in PBS for 15 min at room temperature, washed with phosphate buffered saline and blocked with blocking buffer containing 0.2% TritonX100 and 2% BSA. anti-H3K14ac antibody (Cell Signaling Technologies) in blocking buffer was added and incubated overnight at 4 °C. After washing, a secondary antibody labelled with AlexaFluor 488 dye (ThermoFisher) and Hoechst 33342 (1 μg/ml, Life Technologies) was added for 2 h incubation at room temperature. Plates were washed and read on a PerkinElmer Opera HCS high-content imaging platform. Using a Columbus image analysis pipeline, individual nuclei were located by Hoechst 33342 stain and the level of H3K14ac was calculated from the Alexa-Fluor-488-related intensity in the same area. The resulting mean intensity per cell was converted to per cent inhibition relative to controls on the same plate and the data fitted against a four-parameter logistic model to determine the IC_{50}.

HBO1–BPRF2 protein production, surface plasmon resonance and structural biology

HBO1–BPRF2 protein was produced as previously described[46]. Surface plasmon resonance (SPR) for WM-3835 was done as previously described[25]. Crystals were grown at the CSIRO C3 crystallization centre in SD2 sitting-drop plates at 20 °C, with equal volumes of protein and crystallant (200 nl plus 200-nl drops) with the reservoir consisting of 244 mM diammonium tartrate and 20% PEG 3350. Crystals started to form overnight and were collected 3 days later using 20% glycerol as a cryoprotectant. Data were obtained at the MX2 microfocus beamline at the Australian Synchrotron. The space group was found to be H3, and the data and refinement statistics can be found in Extended Data Fig. 6. The data were indexed with DIALS[47] (WM-3835) or XDS[48] (acetyl-CoA), scaled and integrated with Aimless[49], the structure was solved with Phaser[50] using PDB 5GK9 as the initial model, manually refined with Coot[51] and full refinement was done using Phenix.refine[52] (WM-3835) or REFMAC[53] (acetyl-CoA). Crystallization and refinement statistics are shown in Extended Data Fig. 10.

In vitro metabolic stability

The metabolic stability assay was performed by incubating each test compound in liver microsomes at 37 °C and a protein concentration of 0.4 mg/ml. The metabolic reaction was initiated by the addition of either single cofactor (NADPH only), or dual cofactors (NADPH and UDPGA), and quenched at various time points over a 60-min incubation period by the addition of acetonitrile containing diazepam as an internal standard. Control samples (containing no NADPH) were included (and quenched at 2, 30 and 60 min) to monitor for potential degradation in the absence of cofactor. The human liver microsomes used in this experiment were supplied by XenoTech, lot no. 1410230. The mouse liver microsomes used in this experiment were supplied by XenoTech, lot no. 1510256. Microsomal incubations were performed at a substrate concentration of 1 μM.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

The shRNA screen sequencing data have been deposited to the NCBI Sequence Archieve under the accession number GSE120813. Crystal structure data for HBO1–BPRF2 in complex with WM-3835 and acetyl-CoA have been submitted to the PDB under accession numbers 6MAJ (WM-3835) and 6MAK (acetyl-CoA). Source Data are provided for Figs. 1–4. Any other relevant data are available from the corresponding author upon reasonable request.

Code availability

All code used in this study is publically available.

26. Majewski, I. J. et al. Polycomb repressive complex 2 (PRC2) restricts hematopoietic stem cell activity. *PLoS Biol.* **6**, e93 (2008).
27. Voss, A. K., Collin, C., Dixon, M. P. & Thomas, T. Moz and retinoic acid coordinately regulate H3K9 acetylation, *Hox* gene expression, and segment identity. *Dev. Cell* **17**, 674–686 (2009).
28. Kinkel, S. A. et al. Jarid2 regulates hematopoietic stem cell function by acting with polycomb repressive complex 2. *Blood* **125**, 1890–1900 (2015).
29. Keniry, A. et al. Setdb1-mediated H3K9 methylation is enriched on the inactive X and plays a role in its epigenetic silencing. *Epigenetics Chromatin* **9**, 16 (2016).
30. Dai, Z. et al. edgeR: a versatile tool for the analysis of shRNA-seq and CRISPR-Cas9 genetic screens. *F1000Res.* **3**, 95 (2014).
31. Shechter, D., Dormann, H. L., Allis, C. D. & Hake, S. B. Extraction, purification and analysis of histones. *Nat. Protocols* **2**, 1445–1457 (2007).
32. Cox, J. & Mann, M. MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nat. Biotechnol.* **26**, 1367–1372 (2008).
33. Brumatti, G. et al. HoxA9 regulated Bcl-2 expression mediates survival of myeloid progenitors and the severity of HoxA9-dependent leukemia. *Oncotarget* **4**, 1933–1947 (2013).
34. Tchasovnikarova, I. A. et al. Epigenetic silencing by the HUSH complex mediates position-effect variegation in human cells. *Science* **348**, 1481–1485 (2015).
35. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* **17**, 3 (2011).
36. Kim, D., Langmead, B. & Salzberg, S. L. HISAT: a fast spliced aligner with low memory requirements. *Nat. Methods* **12**, 357–360 (2015).
37. Lavallée, V. P. et al. The transcriptomic landscape and directed chemical interrogation of MLL-rearranged acute myeloid leukemias. *Nat. Genet.* **47**, 1030–1037 (2015).

38. Ritchie, M. E. et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
39. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. B* **57**, 289–300 (1995).
40. Wu, D. et al. ROAST: rotation gene set tests for complex microarray experiments. *Bioinformatics* **26**, 2176–2182 (2010).
41. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
42. Li, H. et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).
43. Orlando, D. A. et al. Quantitative ChIP-seq normalization reveals global modulation of the epigenome. *Cell Reports* **9**, 1163–1170 (2014).
44. Ramírez, F. et al. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, W160–W165 (2016).
45. Robinson, J. T. et al. Integrative Genomics Viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
46. Tao, Y., Zhong, C., Zhu, J., Xu, S. & Ding, J. Structural and mechanistic insights into regulation of HBO1 histone acetyltransferase activity by BRPF2. *Nucleic Acids Res.* **45**, 5707–5719 (2017).
47. Waterman, D. G. et al. Diffraction-geometry refinement in the DIALS framework. *Acta Crystallogr. D* **72**, 558–575 (2016).
48. Kabsch, W. Xds. *Acta Crystallogr. D* **66**, 125–132 (2010).
49. Evans, P. R. & Murshudov, G. N. How good are my data and what is the resolution? *Acta Crystallogr. D* **69**, 1204–1214 (2013).
50. McCoy, A. J. et al. Phaser crystallographic software. *J. Appl. Crystallogr.* **40**, 658–674 (2007).
51. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
52. Afonine, P. V. et al. Towards automated crystallographic structure refinement with phenix. refine. *Acta Crystallogr. D* **68**, 352–367 (2012).
53. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
54. Gilan, O. et al. Functional interdependence of BRD4 and DOT1L in MLL leukemia. *Nat. Struct. Mol. Biol.* **23**, 673–681 (2016).
55. The Cancer Genome Atlas Research Network. Genomic and epigenomic landscapes of adult de novo acute myeloid leukemia. *N. Engl. J. Med.* **368**, 2059–2074 (2013).
56. Bagger, F. O. et al. BloodSpot: a database of gene expression profiles and transcriptional programs for healthy and malignant haematopoiesis. *Nucleic Acids Res.* **44**, D917–D924 (2016).

Acknowledgements We thank C. Lovitt, J. Wingerd, S. Jackson and E. Allan for their technical contributions to this project. The work in the Dawson, Blewitt and Burns laboratories was supported by the Cancer Council Victoria Venture Grant Scheme, and Dawson laboratory work was supported by project grant funding from the National Health and Medical Research Council of Australia (1085015). We thank the following funders for fellowship and grant support: Leukaemia Foundation Australia senior fellowship, Cancer Council Victoria Dunlop Fellopship and Howard Hughes Medical Institute international research scholarship (M.A.D.); Victoria Cancer Agency early-career (L.M.) and mid-career (E.Y.N.L.) fellowships; CSL Centenary fellowship (S.-J.D.), Snowdome Foundation (P.Y.), Maddie Riewoldt's Vision Foundation (Y.-C.C.), Bellberry-Viertel Senior Medical Research Fellowship (M.E.B), Novo Nordisk Foundation Hallas Møller Fellowship NNF14OC0008541 (C.C.), National Health and Medical Research Council of Australia through project grants 1081421 (J.B.B. and T.T.), 575558, 1084248 (A.K.V. and T.T.), research fellowship 1081421 (A.K.V.) and postgraduate scholarship (K.-L.C.). Salary support for M.Z., M.d.S., H.F., C.C., P.S.K., P.A.S., I.P.S. and B.J.M. was provided by the Cancer Therapeutics CRC, funded through the Australian Government's Cooperative Research Centre programme. The Novo Nordisk Foundation Center for Protein Research is supported financially by the Novo Nordisk Foundation (grant agreement NNF14CC0001). This work was made possible through the Victorian State Government Operation Infrastructure Support and Australian National Health and the Medical Research Council Research Institute Infrastructure Support Scheme. We thank the Australian Synchrotron and beamline scientists for help with data collection; this research was undertaken in part using the MX2 beamline at the Australian Synchrotron and made use of the ACRF detector.

Author contributions M.A.D. developed the overall concept behind the study and supervised the project with important advice from L.M., M.E.B., C.J.B., A.K.V., T.T., C.C., F.G. and B.J.M. The manuscript was written by L.M., B.J.M. and M.A.D., with input from all the authors. L.M. performed the majority of the experiments with help from J.A., E.Y.N.L., Y.-C.C., C.-F.W., M.M.Y., P.Y., K.-L.C., K.K., M.S.B., M.L.B., J.L., Y.Y., R.A.B., T.T., H.F., M.d.S., N.N., P.S.K., M.Z., V.M.A. and C.C.-P. Critical reagents and interpretation of the research findings were provided by A.K.V., F.G., S.-J.D., P.A.S., C.C.-P., T.S.P., J.B.B., T.T., O. Dovey, G.S.V., C.C. and I.P.S. The shRNA library was designed and built by L.J.G., T.W., J.L., J.C. and M.E.B. A.H. and C.C. performed mass spectrometry analysis of histone acetylation sites. Protein production, SPR and structural studies were done by S.N., R.S., M.L.D., O. Dolezal, J.N., B.R. and T.S.P. WM-3835 was developed by J.B.B., D.J.L. and N.N. with analytical support by Y.S.

Competing interests M.A.D. has been a member of advisory boards for CTX CRC, Storm Therapeutics, Celgene and Cambridge Epigenetix. The Dawson, Thomas and Voss laboratories receive research funding from CTx CRC. The remaining authors declare no competing financial interests.

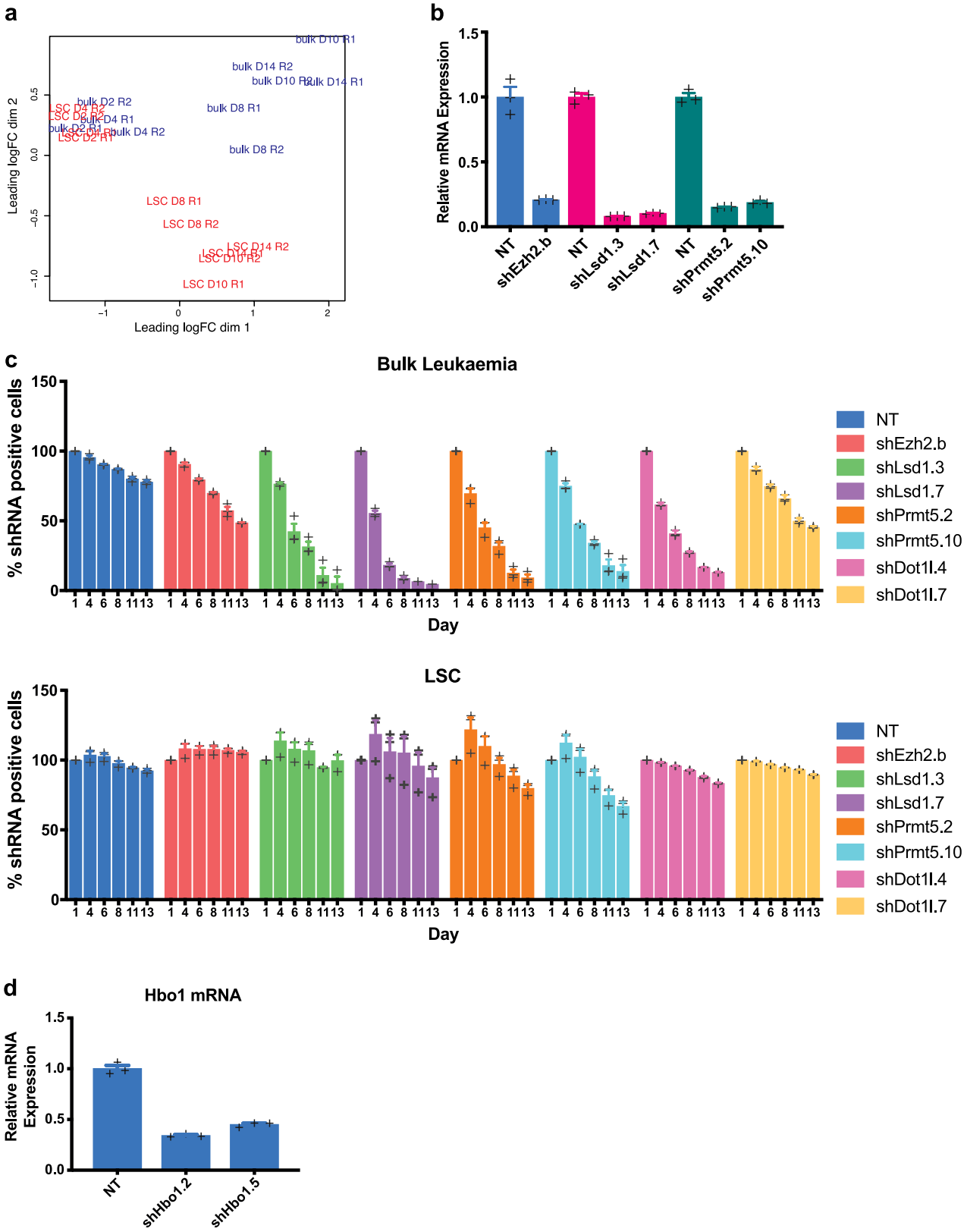
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1835-6.

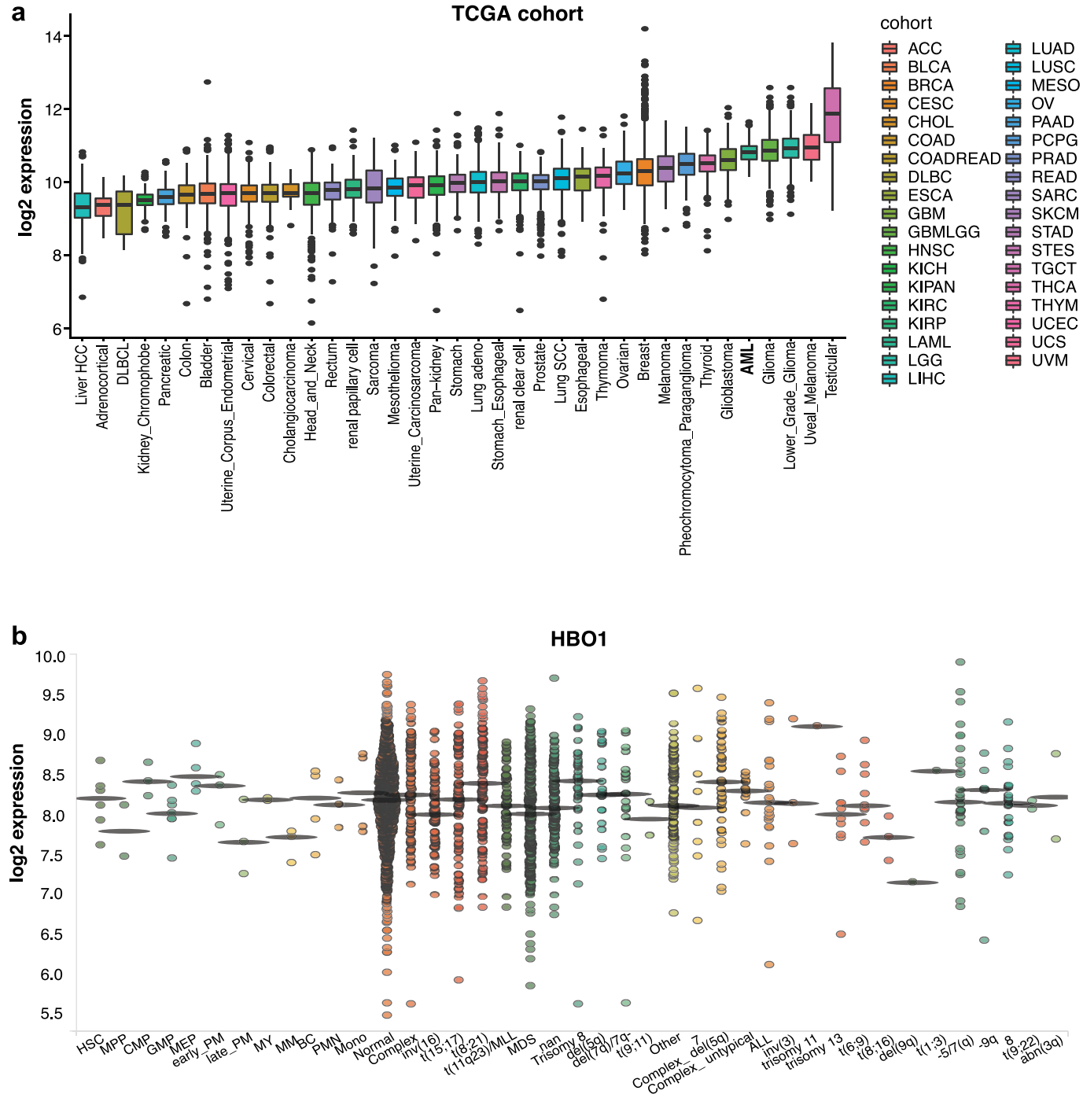
Correspondence and requests for materials should be addressed to M.A.D.

Peer review information *Nature* thanks Stefan Knapp and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

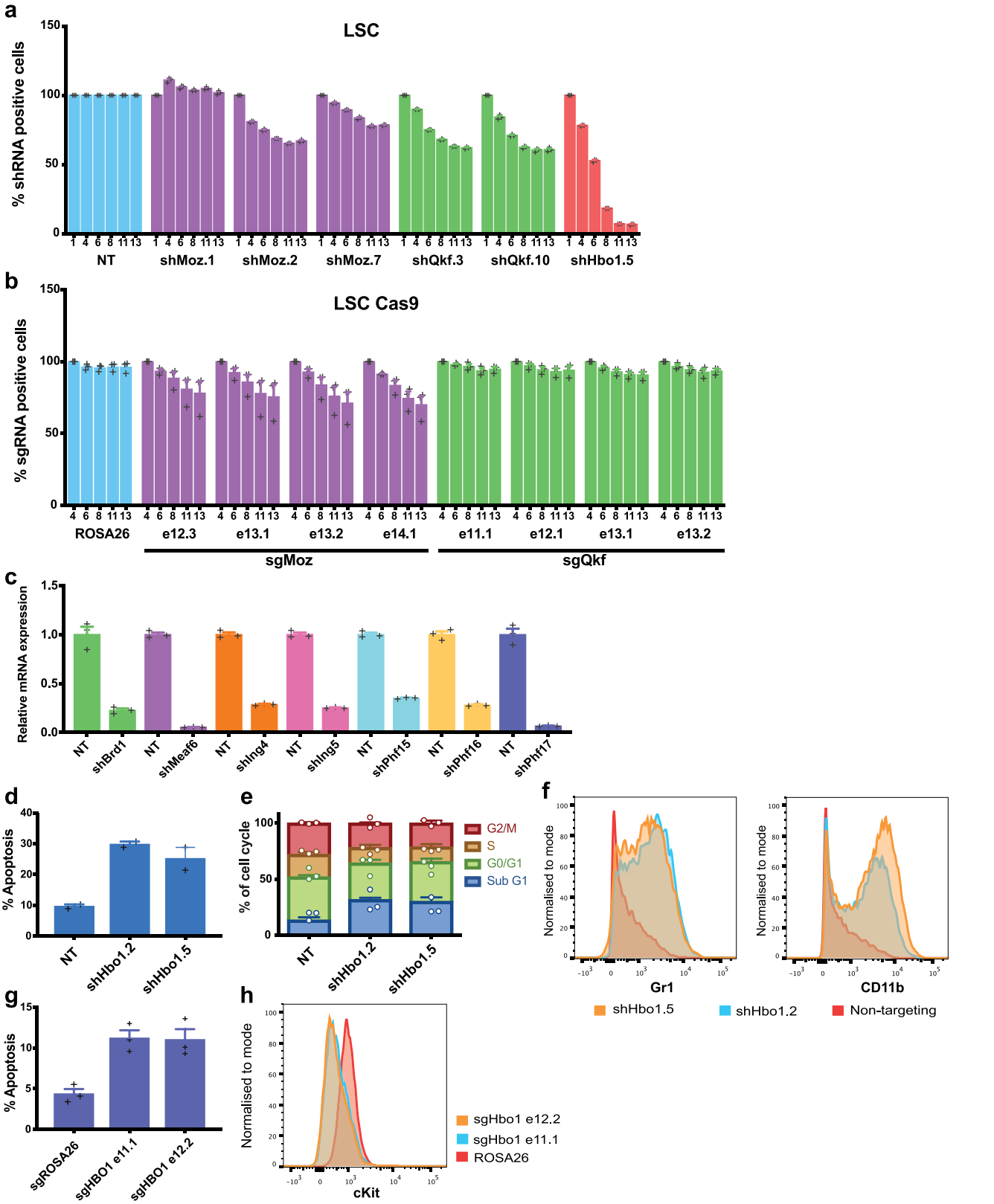
Reprints and permissions information is available at http://www.nature.com/reprints.



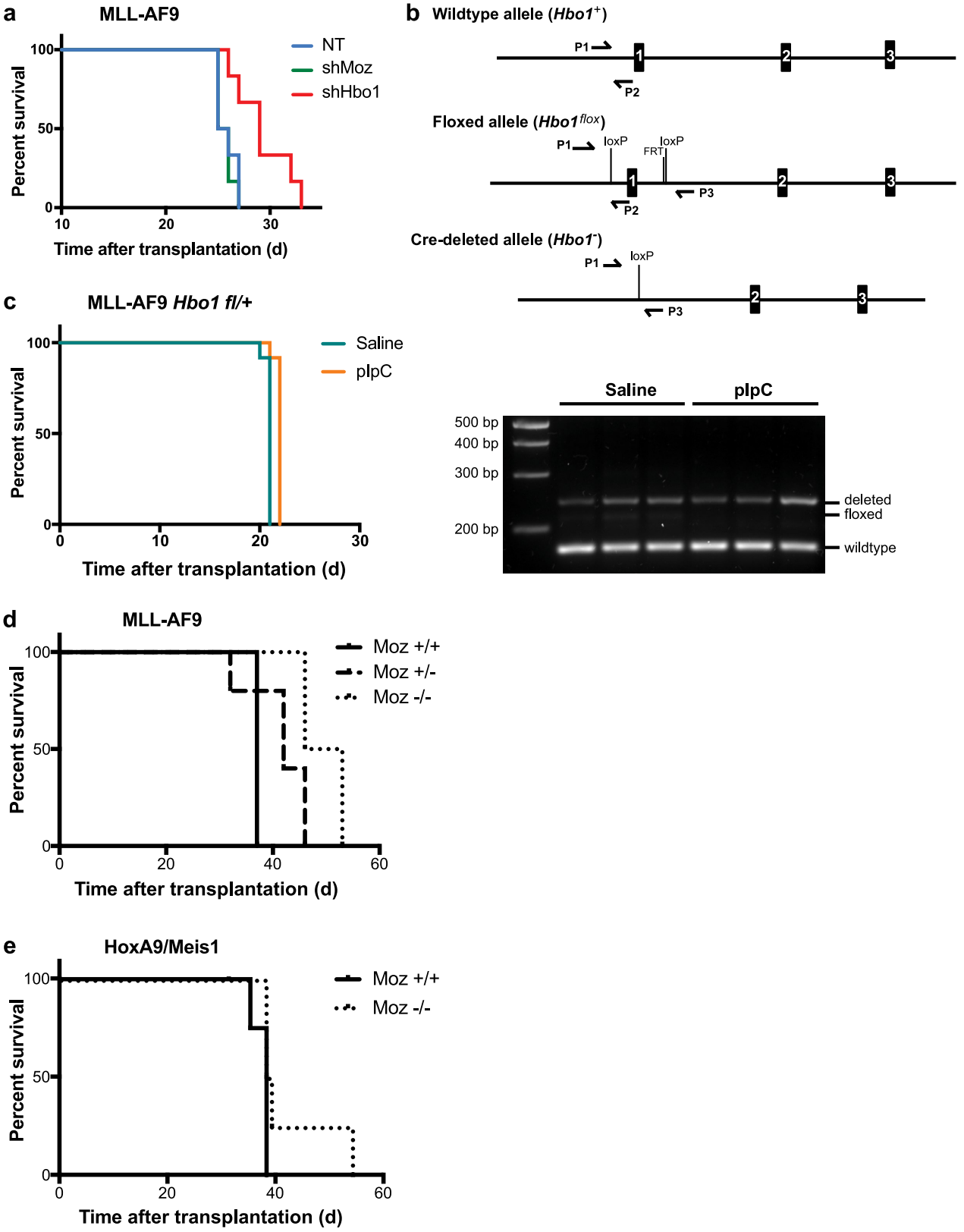
Extended Data Fig. 1 | HBO1 is an AML-specific dependency. a, RNA interference screen multidimensional scaling plot of shRNA sequences from bulk leukaemic blasts (bulk) and LSCs over 14 days of screening. Screens were performed in duplicate (R1 and R2). **b**, Expression of *Ezh2*, *Lsd1* or *Prmt5* mRNA, relative to non-targeting shRNA in LSCs that express shRNAs. $n = 3$. Mean + s.e.m. **c**, Negative-selection competition assays in bulk leukaemic blasts and LSCs that express shRNAs against *Ezh2*, *Lsd1*, *Prmt5* or *Dot1l*. $n = 3$. Mean + s.e.m. Validation of the shRNAs against *Dot1l* have previously been published[54]. **d**. Expression of *Hbo1* mRNA, relative to non-targeting shRNA, in LSCs that express shRNAs. $n = 3$. Mean + s.e.m.



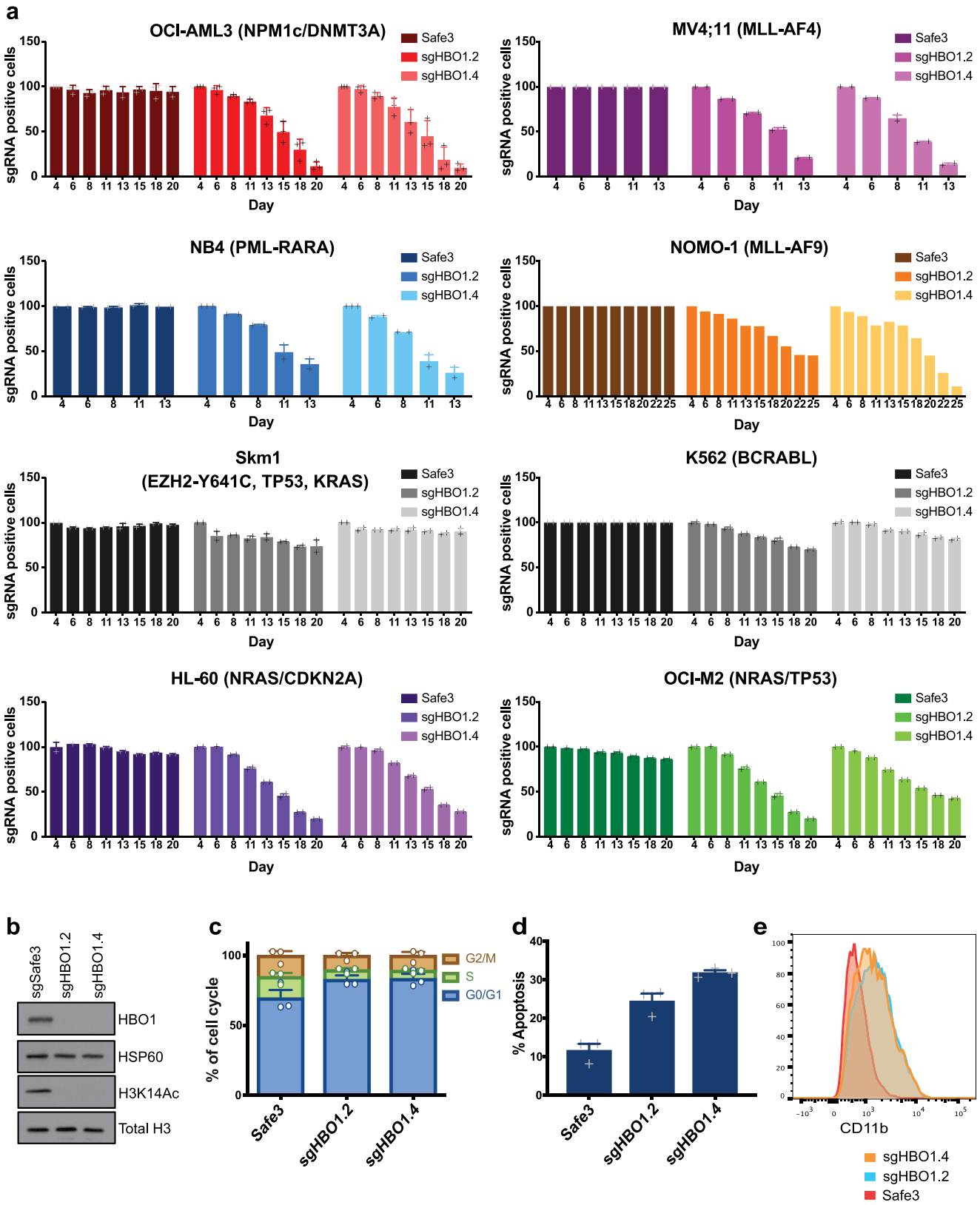
Extended Data Fig. 2 | Expression levels of HBO1. a. Box plot of HBO1 expression levels in cancer types from The Cancer Genome Atlas (TCGA)[55]. The upper limit, centre and lower limit of each box denotes the upper quartile, median and lower quartile of the data, respectively. Cohort abbreviations refer to TCGA study abbreviations (https://gdc.cancer.gov/resources-tcga-users/tcga-code-tables/tcga-study-abbreviations). **b**, Dot plot of levels of HBO1 expression in normal haematopoietic and AML cells from BloodSpot[56]; horizontal line indicates mean expression.



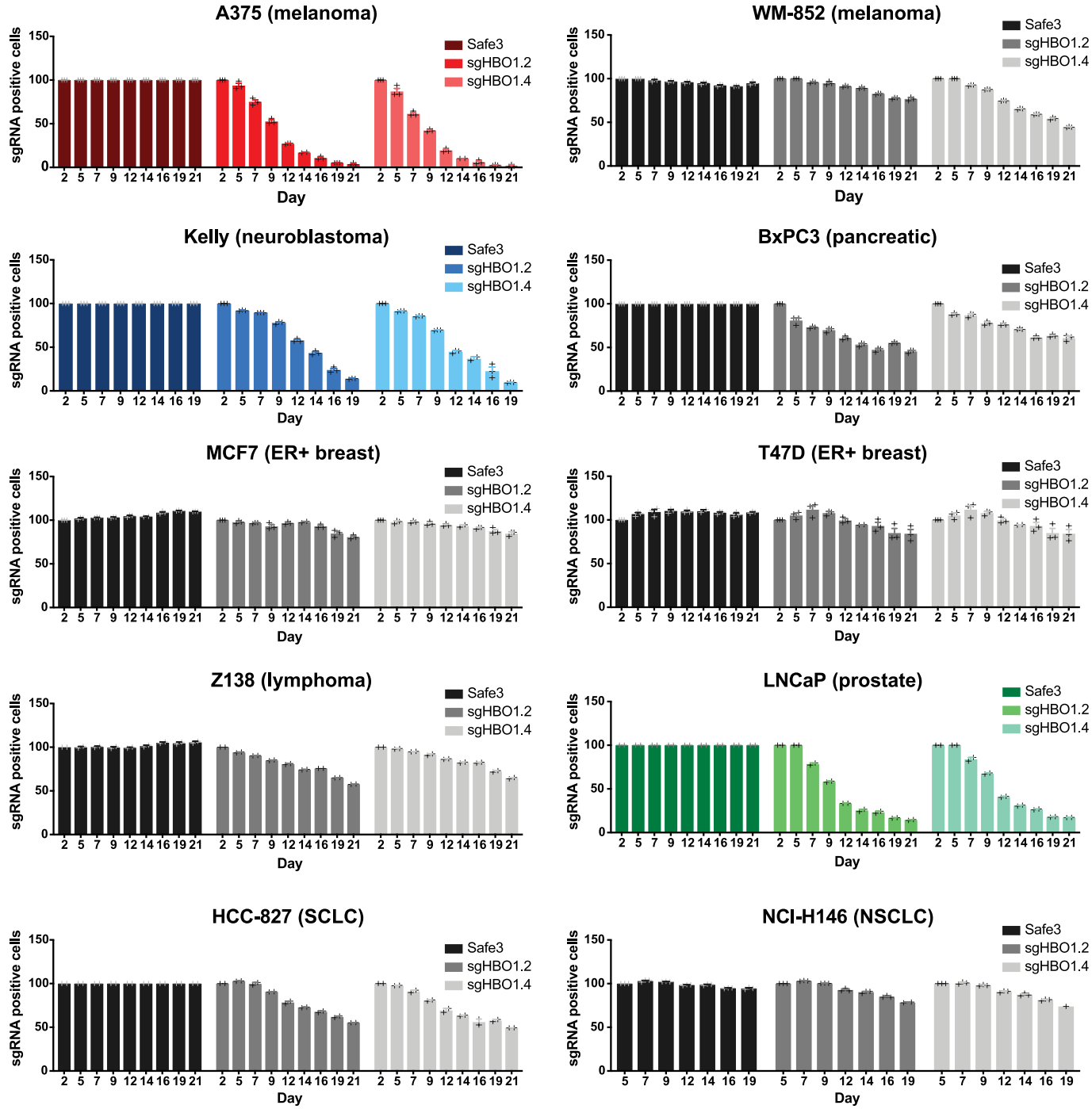
Extended Data Fig. 3 | HBO1 depletion increases apoptosis, cell-cycle arrest and myeloid differentiation in mouse LSCs. a, Negative-selection competition assays in LSCs that express shRNAs against *Moz*, *Qkf* or *Hbo1*. $n = 3$. Mean + s.e.m. **b**, Negative-selection competition assays in LSCs that express Cas9, transduced with sgRNAs targeting *Moz* or *Qkf*. $n = 3$. Mean + s.e.m. **c**, Expression of members of the HBO1 complex, relative to non-targeting shRNA in LSCs that express shRNAs targeting individual members of the complex. $n = 3$. Mean + s.e.m. **d**, Apoptosis of LSCs that express shRNAs targeting *Hbo1*. $n = 3$. Mean + s.e.m. **e**, Cell-cycle profile of LSCs that express shRNAs targeting *Hbo1*. $n = 3$. Mean + s.e.m. **f**, Surface expression of GR1 and CD11b in LSCs that express shRNAs against *Hbo1*. $n = 3$. **g**. Apoptosis of LSCs that express Cas9 and sgRNAs targeting *Hbo1*. $n = 3$. Mean + s.e.m. **h**, Surface expression of KIT (also known as CD117) in LSCs that express sgRNAs targeting *Hbo1*. Representative of $n = 2$ biological replicates.



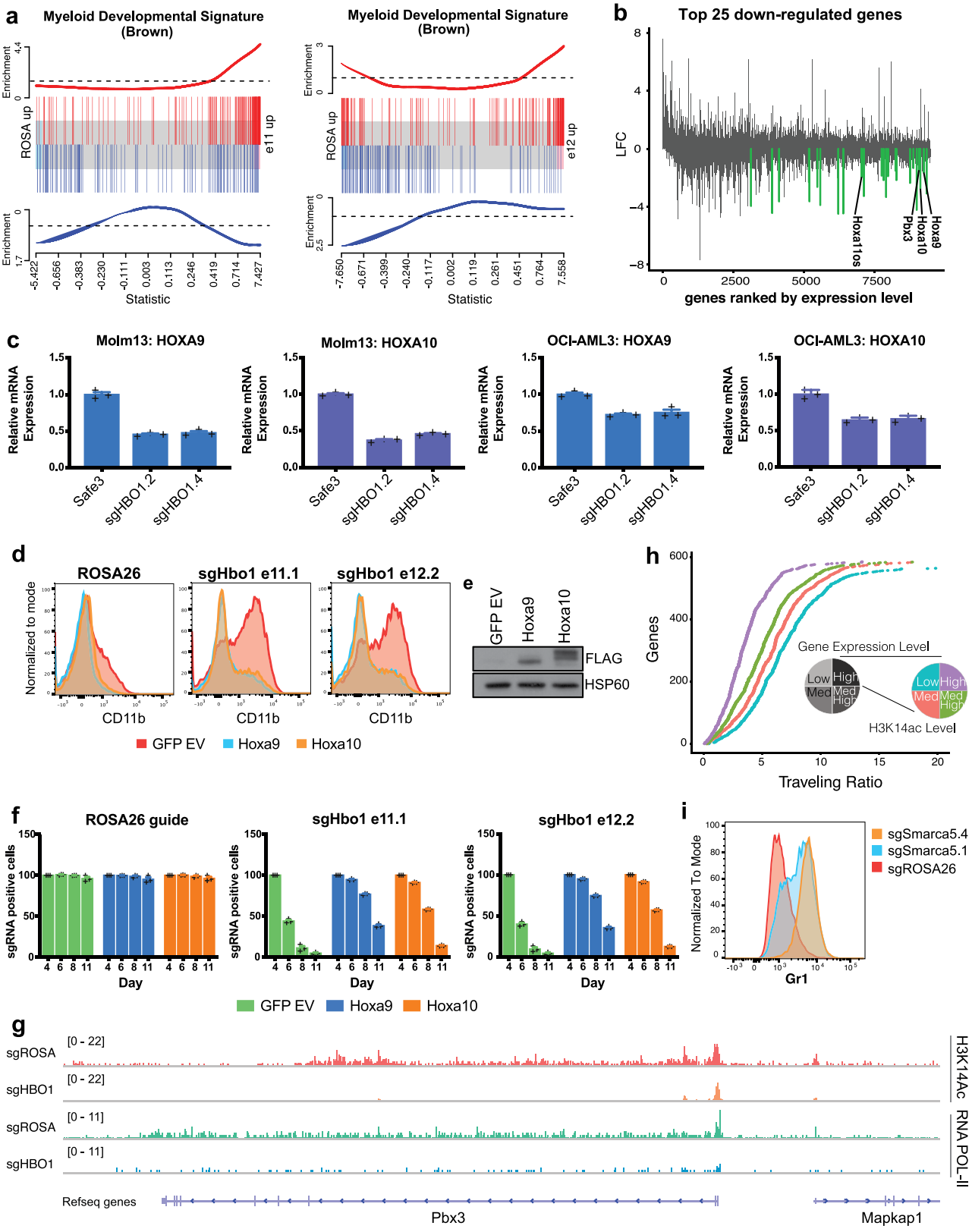
Extended Data Fig. 4 | In vivo depletion of HBO1 increases disease latency. **a**, Kaplan–Meier curves of NSG mice transplanted with quinary MLL–AF9 leukaemic cells that express shRNAs targeting *Hbo1* or *Moz*. $n = 6$ mice per group. **b**, Schematic of wild-type and *Hbo1*-mutant alleles with numbered black boxes representing exons. Genotyping primers are indicated (adapted from ref. [10]). **c**, Kaplan–Meier curves of C57BL/6 mice injected with 10^6 $Hbo1^{flox/+}$ *Mx1-cre* primary MLL–AF9 leukaemic cells. Mice were treated with saline or pIpC. $n = 12$ mice per group. Right, representative genotyping of recipient bone marrow at end point. **d**, Kaplan–Meier curves of C57BL/6 mice injected with $Moz^{+/+}$, $Moz^{+/-}$ or $Moz^{-/-}$ MLL–AF9 leukaemic cells. $n = 5$ mice per group. **e**, Kaplan–Meier curves of BALB/c mice injected with $Moz^{+/+}$ or $Moz^{-/-}$ HOXA9 and MEIS1 leukaemic cells. $n = 5$ mice per group.



Extended Data Fig. 5 | HBO1 is a dependency in various subtypes of AML. **a**, Negative-selection competition assays in human AML cell lines that express Cas9 and sgRNAs targeting *HBO1*. $n = 3$. Mean + s.e.m. Driver mutations are in parentheses. **b**, Immunoblot of HBO1 and H3K14ac in OCI-AML3 cells that express Cas9 and sgRNAs targeting *HBO1*. $n = 3$. Mean + s.e.m. **c**, Cell-cycle profile of OCI-AML3 cells that express Cas9 and sgRNAs targeting *HBO1*. $n = 3$. Mean + s.e.m. **d**, Apoptosis of OCI-AML3 cells that express Cas9 and sgRNAs targeting *HBO1*. $n = 3$. Mean + s.e.m. **e**, Surface expression of CD11b in OCI-AML3 cells that express Cas9 and sgRNAs targeting *HBO1*. $n = 3$.



Extended Data Fig. 6 | HBO1 dependency in other cancers. Negative-selection competition assays in human cancer cell lines that express Cas9 and sgRNAs targeting *HBO1*. $n = 3$. Mean + s.e.m. Cancer type is in parentheses.



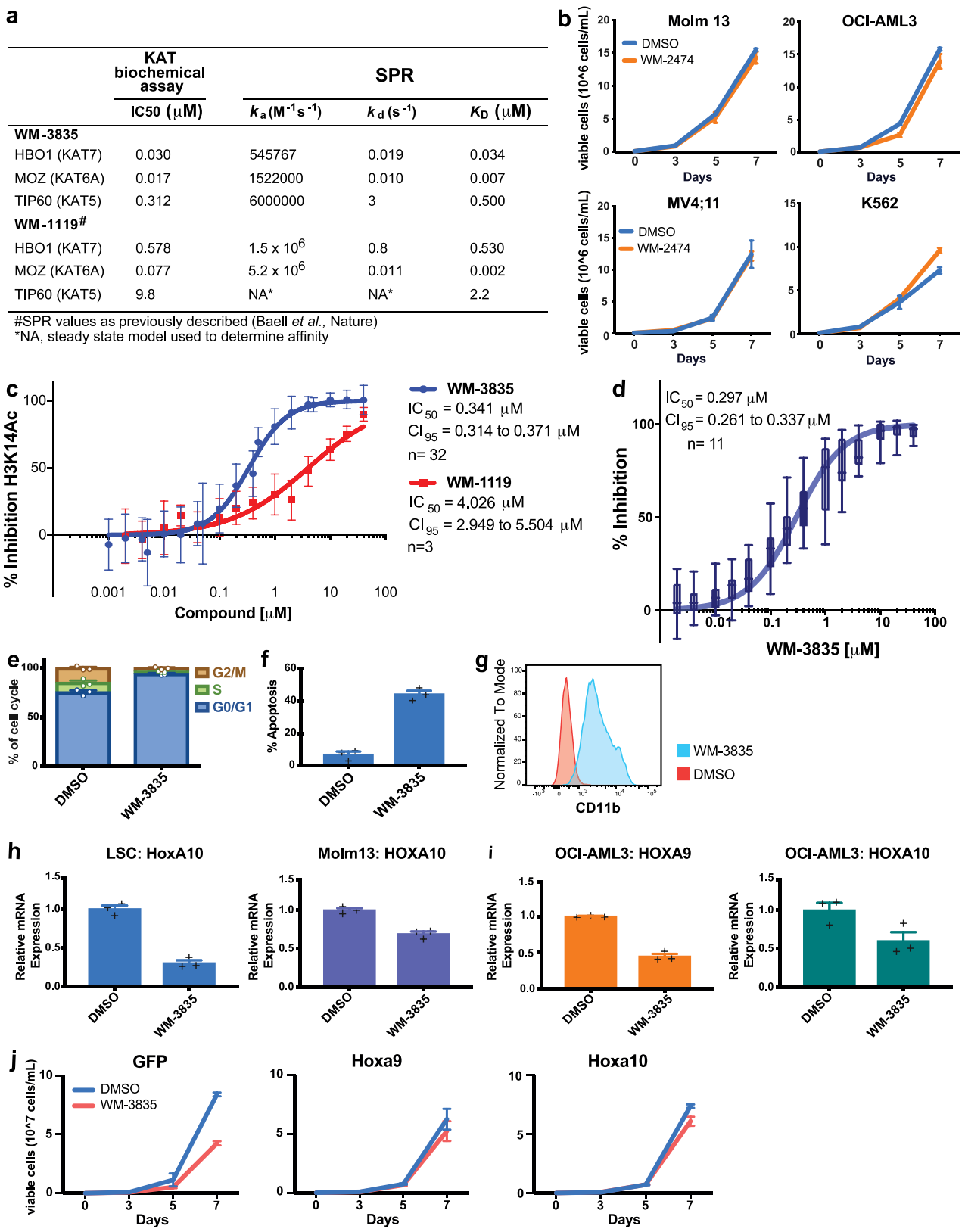
Extended Data Fig. 7 | See next page for caption.

Extended Data Fig. 7 | HBO1 depletion increases myeloid signature and decreases global H3K14ac. a, Barcode plot evaluating changes in myeloid development signature after HBO1 depletion with *Hbo1* e11.1 and *Hbo1* e12.2 sgRNAs in LSCs. $n = 3$. **b**, Bar plot of changes in gene expression after *Hbo1* deletion in LSCs, ranked by expression levels. Green bars show the top 25 most-downregulated genes after *Hbo1* deletion. **c**, *HOXA9* and *HOXA10* mRNA expression in Molm13 and OCI-AML3 cells that express Cas9 and sgRNA targeting *HBO1*. $n = 3$. Mean ± s.e.m. **d**, Surface expression of CD11b in LSCs that overexpress *Hoxa9* or *Hoxa10*, and sgRNAs targeting *Hbo1*. **e**, Immunoblot of overexpressed HOXA9 or HOXA10 in LSCs that express Cas9. Representative of $n = 3$ biological replicates. **f**, *Hoxa9*- and *Hoxa10*-rescue negative-selection competition assays. Representative of $n = 3$ biological replicates. **g**, ChIP–seq profiles of H3K14ac and RNA Pol-II at the *Pbx3* locus in LSCs that express Cas9 and an sgRNA targeting *Hbo1*. Representative of $n = 3$ biological replicates. **h**, RNA Pol-II travelling ratio distribution for highly expressed genes, divided according to H3K14ac levels from ChIP–seq. **i**, Surface expression of GR1 in LSCs that overexpress sgRNAs targeting *Smarca5*. Representative of $n = 3$ biological replicates.

a

	KAT biochemical assay	SPR		
	IC50 (μM)	k_a ($M^{-1}s^{-1}$)	k_d (s^{-1})	K_D (μM)
WM-3835				
HBO1 (KAT7)	0.030	545767	0.019	0.034
MOZ (KAT6A)	0.017	1522000	0.010	0.007
TIP60 (KAT5)	0.312	6000000	3	0.500
WM-1119#				
HBO1 (KAT7)	0.578	1.5 x 10^6	0.8	0.530
MOZ (KAT6A)	0.077	5.2 x 10^6	0.011	0.002
TIP60 (KAT5)	9.8	NA*	NA*	2.2

#SPR values as previously described (Baell *et al.*, Nature)
*NA, steady state model used to determine affinity

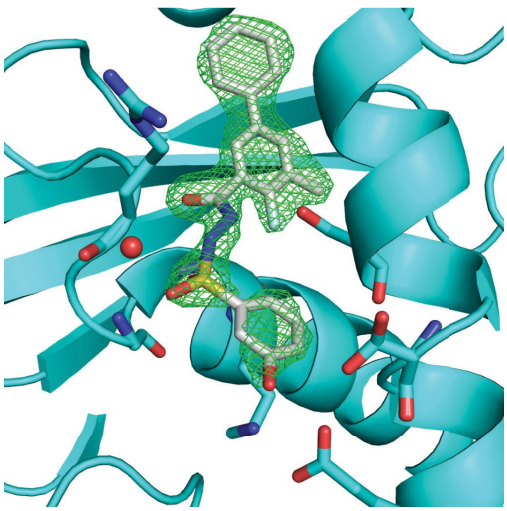


Extended Data Fig. 8 | WM-3835 inhibits cell growth and HOXA expression in AML. a, KAT biochemical and SPR values for WM-3835 compared to WM-1119. Biochemical assay was done at 1 μM acetyl-CoA, the K_m of HBO1. **b**, Proliferation assays of human AML cells treated with 1 μM WM-2474. $n = 3$. Mean + s.e.m. **c**, Cellular H3K14ac biomarker assay dose–response curves for WM-3835 (blue) and WM-1119 (red). $n = 6$. Mean ± s.e.m. **d**, Growth inhibition assays of the MLL–AF9 AML cell line Molm13 treated with WM-3835 at the doses indicated. Boxes represent minimum and maximum values. $n = 11$. Mean ± s.e.m. **e**, Cell-cycle profile of OCI-AML3 cells treated with WM-3835 or vehicle. $n = 3$. Mean + s.e.m. **f**, Apoptosis of OCI-AML3 cells treated with WM-3835 or vehicle. $n = 3$. Mean + s.e.m. **g**, Surface expression of CD11b in OCI-AML3 cells treated with WM-3835 or vehicle. $n = 3$. **h**, *HOXA10* mRNA expression in LSCs and Molm13 cells treated with WM-3835 or vehicle. $n = 3$. Mean + s.e.m. **i**, *HOXA9* and *HOXA10* mRNA expression in OCI-AML3 cells treated with WM-3835. $n = 3$. Mean + s.e.m. **j**, *Hoxa9* and *Hoxa10* rescue proliferation assays with 1 μM WM-3835 in LSCs. $n = 3$. Mean ± s.e.m.

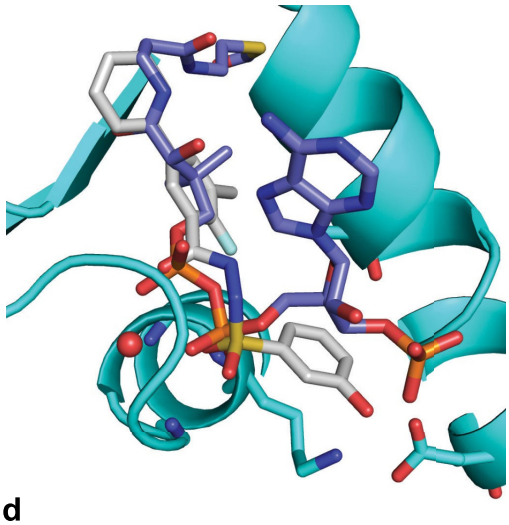
a

	WM-3835	acetyl-CoA
Data collection		
Space group	H3	I2
Cell dimensions		
a, b, c (Å)	111.6, 111.6, 73.8	87.6, 39.0, 107.8
a, b, g (°)	90, 90, 120	90, 99.6, 90
Resolution (Å)	2.14 (2.20 - 2.14)	2.13 (2.19 - 2.13)
R_{merge}	0.198 (1.292)	0.070 (0.748)
R_{pim}	0.095 (0.614)	0.029 (0.310)
I / sI	5.7 (1.9)	13.1 (2.6)
CC1/2	0.987 (0.583)	0.998 (0.882)
Completeness (%)	100 (100)	98.0 (94.2)
Redundancy	5.3 (5.3)	6.9 (6.6)
Refinement		
Resolution (Å)	58.6 - 2.14	43.2 - 2.13
Unique reflections	18,829	18,965
R_{work} / R_{free}	19.5 / 23.7	21.5 / 26.4
No. atoms	2553	2488
Protein (HBO1)	2225	2207
Protein (BRPF2)	143	206
Inhibitor	28	51
Water	157	23
B-factors (Å^2)		
Protein (HBO1)	40.0	59.7
Protein (BRPF2)	78.5	75.5
Inhibitor	19.9	61.0
Water	43.6	49.0
R.m.s deviations		
Bond lengths (Å)	0.004	0.008
Bond angles (°)	0.733	1.189
Note: Values in parentheses are for the highest resolution shell.		

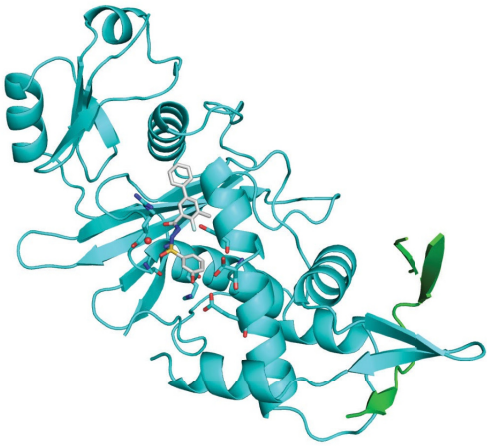
b



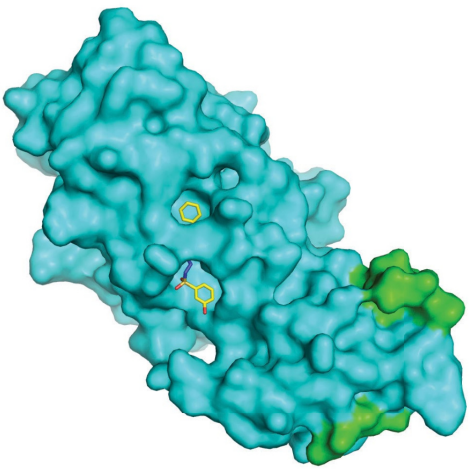
c



d

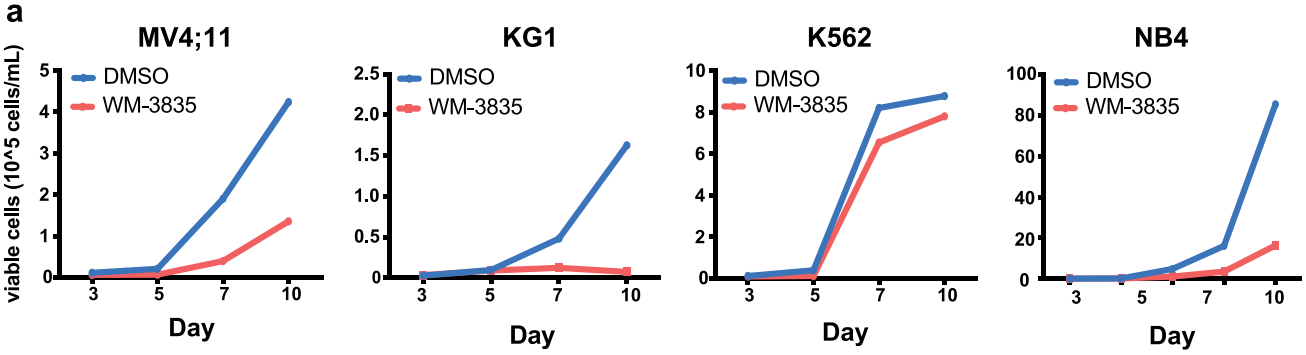


e

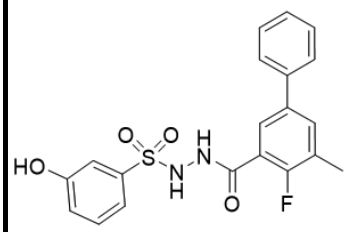


Extended Data Fig. 9 | HBO1 crystal structure. a, Data collection and refinement statistics of WM-3835 HBO1–BRPF2 co-crystal structure. **b**, WM-3835 binding site in HBO1–BRPF2. WM-3835 is shown in silver with element colouring and the omit electron density map, contoured to 3σ, is shown in green. **c**, Overlay of WM-3835 and acetyl-CoA (purple with element colouring), showing that WM-3835 binds in the acetyl-CoA binding site of HBO1. **d**, Ribbon diagram of HBO1–BRPF2 showing WM-3835 bound to the acetyl-CoA binding site. **e**, Space-filling model showing WM-3835 (yellow with element colouring) in the acetyl-CoA binding pocket of HBO1–BRPF2.

a



b

Compound (Batch)	Structure	Cofactor	Microsome Species	$T_{1/2}$ (min)	$CL_{int,\ in\ vitro}$ (μL/min/mg protein)
WM-3835		Single (NADPH)	Human	17	104
			Mouse	5	347
		Dual (NADPH + UDPGA)	Human	8	230
			Mouse	4	402

c

Treatment	Mouse Number	Total drug concentration (μM)	Free drug concentration (μM)#	Glucuronide peak area (M+176)*
WM-3835 (100 mg/kg, BID)	1	1.28	0.0017	201576
	2	1.75	0.0024	213458
	3	2.14	0.0029	218945
	4	2.25	0.0031	214612
mean		**1.86**	**0.0026**	
SD		**0.44**	**0.0006**	

Corrected for mouse plasma binding (fu 0.0014)

* Elution of the putative glucuronide monitored using predicted MRM of 576.98 > 400.98 (cone voltage 35 eV and collision energy of 20 eV).

Extended Data Fig. 10 | See next page for caption.

Extended Data Fig. 10 | High in vitro metabolism and poor in vivo oral exposure of WM-3835. a, Proliferation assays of human AML cell lines treated with 1 μM WM-3835. **b**, WM-3835 demonstrates high clearance in both human and mouse liver microsome assays. The use of dual cofactors (UDPGA and NADPH) results in an increased rate of clearance in human liver microsomes, which is consistent with glucuronidation having a role in the clearance of this compound. **c**, BALB/c female mice were dosed with WM-3835 at 100 mg kg^{-1} taken by mouth twice a day, formulated in 20% PEG400/10% solutol or vehicle. Four hours after the third dose, blood samples were collected. An average total drug concentration of 1,860 nM was observed. The free drug level was determined to be 2.6 nM after accounting for mouse plasma protein binding (fraction unbound 0.0014). This free drug level was considered too low to affect the acetylation of H3 at K14, on the basis of the in vitro H3K14ac cellular biomarker data (Fig. 4). An additional chromatographic peak eluting earlier than WM-3835 was detected in the plasma samples from the treatment group. Subsequent analyses using predicted multiple-reaction monitoring and accurate mass measurement indicated that this peak is likely to be a glucuronide conjugate of WM-3835, consistent with the in vitro metabolism data.

nature research

Corresponding author(s): Mark Dawson

Last updated by author(s): Oct 9, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☐	☒ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒ A description of all covariates tested
☐	☒ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Flow cytometry data was collected using LSR II, FACSVERSE, FACSFortessa X20 (BD Biosciences); Sequencing data was collected using Next-Seq500 (Illumina); Quantative PCR data was collected using StepOnePlus Real-Time PCR system (Applied Biosystems)
Data analysis	BD FACSuite (BD Biosciences), FlowJo V10, and FlowLogic 7 (Inivai) software were used to collect and analyze flow cytometry data. Limma (v3.34.9) and edgeR (v3.20.9) was used to analyse RNA sequencing data. Ngsplot (v2.61) and macs2 (v2.1.1) were used to analyse ATAC and ChIP sequencing data. StepOnePlus (Applied Biosystems) software was used to analyse quantitative PCR data. Statistical analysis was performed using R or Prism 7 (GraphPad), MaxQuant was used to process LC-MS/MS data.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

The sequencing data that support the findings of this study has been deposited into the sequence read archive, which is hosted by the National Centre for Biotechnology Information.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	The number of mice assigned to each treatment arm was selected to provide sufficient statistical power to discern significant differences.
Data exclusions	Mice were excluded from the Kaper-Meier curves if they had less than 20% detectable leukemic cells in the bone marrow or spleen at the time of death. The only other data points excluded were those that were clear outliers in technical replicates of proliferation assays or qRT-PCR.
Replication	All in vitro functional and phenotypic experiments were performed in at least biological triplicate to ensure reproducibility. All experiments were able to be reliably reproduced.
Randomization	Allocation of mice to different groups was random.
Blinding	Experimenters were not blinded to the experiment. However, all mouse experiments were performed by the core facility staff who have no direct involvement in the project and were blinded to the genotype of the mice or the shRNA that was used in the experiment.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☐	☒ Antibodies
☐	☒ Eukaryotic cell lines
☒	☐ Palaeontology
☐	☒ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☐	☒ ChIP-seq
☐	☒ Flow cytometry
☒	☐ MRI-based neuroimaging

Antibodies

Antibodies used	rabbit anti-acetyl-histone H3 (Lys 14) (D4B9, Cell Signalling Technology, Lot# 4) WB: 1:1000, ChIP-seq: 10 ug, mouse anti-histone H3K14Ac (13HH3-1A5, Active Motif, Lot# 10715003/11709001) WB: 1:1000, rabbit anti-histone H3 (acetyl K27) (ab4729, Abcam, Lot# GR3205526-1/GR3211959-1) WB: 1:5000, rabbit anti-KAT7/Hbo1/MYST2 (ab70183, Abcam, Lot# GR222365-1) WB: 1:1000, rabbit anti-histone H3 (ab1791, Abcam, Lot# GR234589-1) WB: 1:10000, mouse anti-HSP60 (C10, Santa Cruz, Lot# H1913) WB: 1:10000, rabbit anti-HSP60 (H-300, Santa Cruz, Lot# G0119) WB: 1:10000, mouse anti-FLAG (M2, Sigma, Lot# SLBW3851) WB: 1:2000, mouse anti-RNA polymerase II (CTD4H8, Millipore, Lot# 2958013) ChIP-seq: 10 ug. Gt-anti-Rb IgG (H+L) secondary antibody, HRP conjugate (65-6120, Invitrogen, Lot# TH275335), Goat anti-mouse Ig, Human ads-HRP (1010-05, Southern Biotech, Lot# H0416-T867C).WB secondary antibodies (both anti-Rb and -Ms) were used at dilutions ranging 1:10000 to 1:50000.Flow Cytometry: Alexa Fluor 700 anti-Gr1 (108422, BioLegend, Lot# B199432/B258102) and Brilliant Violet 605 anti-CD11b (101237, Biolegend, Lot# B266535), APC-Cy7 anti-mouse CD117 (c-kit) (313228, Biolegend, Lot# B163721), FITC-Annexin V (640906, Biolegend, Lot# 8311824).
Validation	All of the antibodies used have been extensively utilized in the literature and have been validated previously. H3K14ac and H3K23ac antibodies have been validated using the Epicypher panel. H3K14ac and Hbo1 antibodies were validated for WB in house using knockout cell lines (Cre lox and CRISPR-Cas9). H3K27ac and Total H3 for WB have and RNA Pol-II for ChIP-seq have been validated and previously published by our laboratory (Gilan et al, NSMB, 2016). All flow cytometry antibodies have been previously validated by our laboratory (Fong et al., Nature, 2015).

Eukaryotic cell lines

Policy information about cell lines

Cell line source(s)	LSC and Blast cell lines were generated in house (Fong et al, Nature, 2015) Molm13, MV4;11, K562, HL-60, MCF7, T47D, MDA-MB-231, Z-138, HCC-827, NCI-H146, LnCaP, BxPC3 and A375 cell lines were obtained from ATCC NOMO-1 cells were obtained from JCRB OCI-AML3, OCI-M2, NB4, Skm1 and Kelly cell lines were obtained from DSMZ WM-852 were obtained from ESTDAB
Authentication	STR testing was performed to authenticate cell lines
Mycoplasma contamination	All cell lines used are mycoplasma negative as tested by the Victorian Infectious Diseases References Laboratory
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.

Animals and other organisms

Policy information about studies involving animals; ARRIVE guidelines recommended for reporting animal research

Laboratory animals	C57BL/6 female mice All mice were 6–8 weeks old Further details in Methods.
Wild animals	This study did not involve wild animals
Field-collected samples	No data in this study were field-collected
Ethics oversight	All scientific procedures using animals were carried out in accordance with the Prevention of Cruelty to Animals Act 1986 Act (the Act), associated Regulations and the Australian Code of Practice for the Care and Use of Animals for Scientific Purposes (the Code). Experiments were performed under the approved animal ethics experimentation committee permit number E615

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Human research participants

Policy information about studies involving human research participants

Population characteristics	Human bone marrow samples from patients with acute myeloid leukaemia. Patient characteristics are representative of the general population.
Recruitment	Patients are recruited via a PICF to participate in the project titled Molecular Characterisation of Response, Resistance and Clonal Evolution in Haematological Malignancies. We do not foresee biases or any impact on results due patient this selection criteria.
Ethics oversight	The project has been reviewed and approved by the Peter MacCallum Cancer Centre Human Research Ethics Committtte Reference number: HREC/17/PMCC/69

Note that full information on the approval of the study protocol must also be provided in the manuscript.

ChIP-seq

Data deposition

☒ Confirm that both raw and final processed data have been deposited in a public database such as GEO.

☒ Confirm that you have deposited or provided access to graph files (e.g. BED files) for the called peaks.

Data access links *May remain private before publication.*	*For "Initial submission" or "Revised version" documents, provide reviewer access links. For your "Final submission" document, provide a link to the deposited data.*
Files in database submission	*Provide a list of all files available in the database submission.*
Genome browser session (e.g. UCSC)	*Provide a link to an anonymized genome browser session for "Initial submission" and "Revised version" documents only, to enable peer review. Write "no longer applicable" for "Final submission" documents.*

Methodology

Replicates	*Describe the experimental replicates, specifying number, type and replicate agreement.*

Sequencing depth	*Describe the sequencing depth for each experiment, providing the total number of reads, uniquely mapped reads, length of reads and whether they were paired- or single-end.*
Antibodies	*Describe the antibodies used for the ChIP-seq experiments; as applicable, provide supplier name, catalog number, clone name, and lot number.*
Peak calling parameters	*Specify the command line program and parameters used for read mapping and peak calling, including the ChIP, control and index files used.*
Data quality	*Describe the methods used to ensure data quality in full detail, including how many peaks are at FDR 5% and above 5-fold enrichment.*
Software	*Describe the software used to collect and analyze the ChIP-seq data. For custom code that has been deposited into a community repository, provide accession details.*

Flow Cytometry

Plots

Confirm that:

☒ The axis labels state the marker and fluorochrome used (e.g. CD4-FITC).

☒ The axis scales are clearly visible. Include numbers along axes only for bottom left plot of group (a 'group' is an analysis of identical markers).

☐ All plots are contour plots with outliers or pseudocolor plots.

☐ A numerical value for number of cells or percentage (with statistics) is provided.

Methodology

Sample preparation	Cultured cell lines
Instrument	BD Bioscience BD FACSVERSE BD FACSFortessa
Software	BD Bioscience BD FACSuite software was used to collect and analyse data, and FlowJo
Cell population abundance	*Describe the abundance of the relevant cell populations within post-sort fractions, providing details on the purity of the samples and how it was determined.*
Gating strategy	FSC-A and FSC-H was used to identify single cells, cells were separated by viability using DAPI or PI exclusion and (AnnexinV for apoptosis assay)

☐ Tick this box to confirm that a figure exemplifying the gating strategy is provided in the Supplementary Information.

Article

Structural basis of DNA targeting by a transposon-encoded CRISPR–Cas system

https://doi.org/10.1038/s41586-019-1849-0

Received: 16 July 2019

Accepted: 25 October 2019

Published online: 18 December 2019

Tyler S. Halpin-Healy[1], Sanne E. Klompe[1], Samuel H. Sternberg[1]* & Israel S. Fernández[1]*

Bacteria use adaptive immune systems encoded by CRISPR and Cas genes to maintain genomic integrity when challenged by pathogens and mobile genetic elements[1–3]. Type I CRISPR–Cas systems typically target foreign DNA for degradation via joint action of the ribonucleoprotein complex Cascade and the helicase–nuclease Cas3[4,5], but nuclease-deficient type I systems lacking Cas3 have been repurposed for RNA-guided transposition by bacterial Tn*7*-like transposons[6,7]. How CRISPR- and transposon-associated machineries collaborate during DNA targeting and insertion remains unknown. Here we describe structures of a TniQ–Cascade complex encoded by the *Vibrio cholerae* Tn*6677* transposon using cryo-electron microscopy, revealing the mechanistic basis of this functional coupling. The cryo-electron microscopy maps enabled de novo modelling and refinement of the transposition protein TniQ, which binds to the Cascade complex as a dimer in a head-to-tail configuration, at the interface formed by Cas6 and Cas7 near the 3′ end of the CRISPR RNA (crRNA). The natural Cas8–Cas5 fusion protein binds the 5′ crRNA handle and contacts the TniQ dimer via a flexible insertion domain. A target DNA-bound structure reveals critical interactions necessary for protospacer-adjacent motif recognition and R-loop formation. This work lays the foundation for a structural understanding of how DNA targeting by TniQ–Cascade leads to downstream recruitment of additional transposase proteins, and will guide protein engineering efforts to leverage this system for programmable DNA insertions in genome-engineering applications.

We previously demonstrated that a transposon derived from *V. cholerae* Tn*6677* undergoes programmable transposition in *Escherichia coli* directed by a crRNA, and that this activity requires four transposon- and three CRISPR-associated genes in addition to a CRISPR array[7] (Fig. 1a). Whereas TnsA, TnsB and TnsC exhibit functions that are consistent with their homologues from the related and well-studied cut-and-paste DNA transposon *E. coli* Tn*7*[8], we showed that TniQ, a homologue of *E. coli* TnsD, forms a co-complex with the Cascade ribonucleoprotein complex encoded by the type I-F variant CRISPR–Cas system. This finding suggested an alternative role for TniQ, compared with the role of *E. coli* TnsD in identifying target sites during Tn*7* transposition. We proposed that RNA-guided DNA targeting by Cascade could deliver TniQ to DNA in a manner compatible with downstream transpososome formation, and that TniQ might interact with Cascade near the 3′ end of the crRNA, consistent with RNA-guided DNA insertion occurring around 49 bp downstream of the protospacer-adjacent motif (PAM)-distal edge of the target site. To determine this unambiguously, we purified the *V. cholerae* TniQ–Cascade complex loaded with a native crRNA and determined its structure by cryo-electron microscopy (cryo-EM) (Supplementary Table 1).

Cryo-EM structure of TniQ–Cascade complex

The overall complex adopts a helical architecture with protuberances at both ends (Fig. 1, Extended Data Figs. 1, 2). The global architecture is similar to previously determined structures of Cascade from I-E and I-F systems[9–12] (Extended Data Fig. 3), with the exception of a large mass of additional density attributable to TniQ (see below). Maximum likelihood classification methods implemented in Relion3[13] enabled us to identify marked dynamics in the entire complex, which appears to 'breathe', widening and narrowing the distance between the two protuberances (Extended Data Fig. 1d, Supplementary Video 1). The large subunit encoded by a natural Cas8–Cas5 fusion protein (hereafter referred to simply as Cas8) forms one protuberance and recognizes the 5′ end of the crRNA via base- and backbone-specific contacts (Extended Data Figs. 4, 5a–c, 6a), similar to the canonical roles of Cas8 and Cas5 (Extended Data Fig. 3). Cas8 contains two primary subdomains formed mainly by α-helices and a third domain of approximately 100 residues (residues 277 to 385) that is predicted to form three α-helices but could not be built in our maps owing to its intrinsic flexibility (Fig. 1c). However, low-pass-filtered maps revealed that this flexible domain connects with the TniQ protuberance at the opposite end of the crescent-shaped complex (Extended Data Fig. 2e). Additionally, there seemed to be a loose coupling between the Cas8 flexible domain and overall breathing of the complex, as stronger density for that domain could be observed in the closed state (Extended Data Fig. 1d, Supplementary Video 1).

Six Cas7 subunits protect much of the crRNA by forming a helical filament along its length (Fig. 1b, d), similar to other type I Cascade complexes[9–12] (Extended Data Fig. 3). A 'finger' motif in Cas7 clamps the crRNA at regular intervals, causing every sixth nucleotide (nt) of the 32-nt spacer to flip out while leaving the flanking nucleotides

[1]Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA. *e-mail: shsternberg@gmail.com; isf2106@cumc.columbia.edu

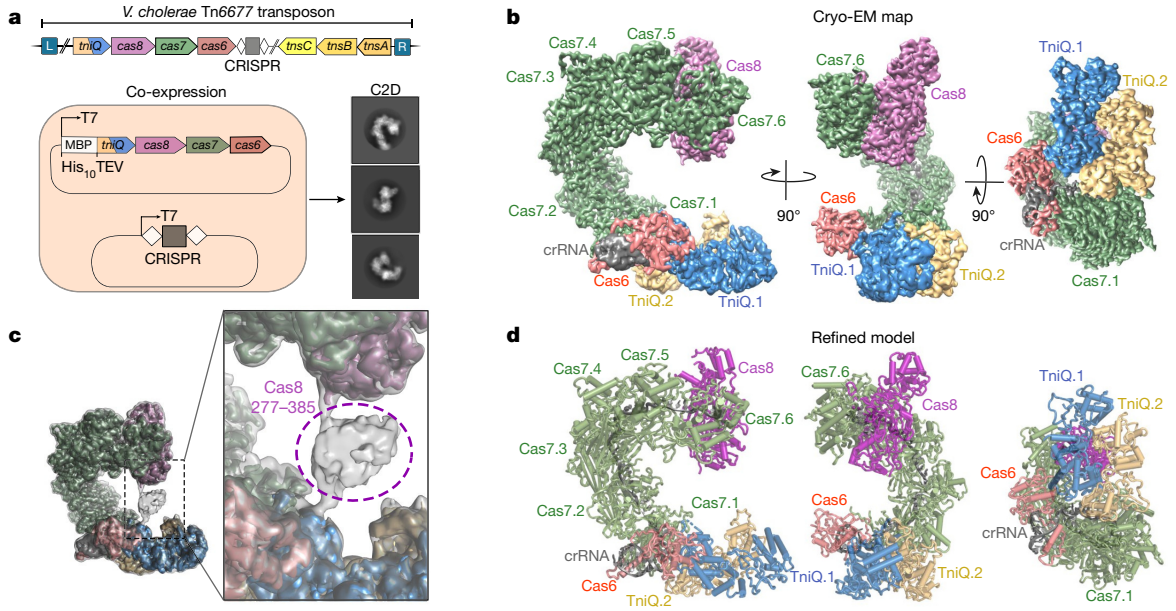


Fig. 1 | Overall architecture of the *V. cholerae* TniQ–Cascade complex. **a**, Genetic architecture of the Tn*6677* transposon (top), and plasmid constructs used to express and purify the TniQ–Cascade complex. Right, selected cryo-EM reference-free two-dimensional class averages in multiple orientations. **b**, Orthogonal views of the cryo-EM map of the TniQ–Cascade complex, showing Cas8 (purple), six Cas7 monomers (green), Cas6 (salmon), crRNA (grey) and TniQ monomers (blue, yellow). The complex adopts a helical architecture with protuberances at both ends. **c**, A flexible domain in Cas8 comprising residues 277–385 (grey) could only be visualized in low-pass-filtered maps. The unsharpened map is shown as semi-transparent, grey map overlaid on the post-processed map segmented and coloured as in **a**. **d**, Refined model for the TniQ–Cascade complex derived from the cryo-EM maps shown in **b**.

available for DNA recognition (Extended Data Figs. 4f, 6). These bases are pre-ordered in short helical segments, with a conserved phenylalanine stacking below the first base of every segment. Cas7.1, the monomer furthest away from Cas8, interacts with Cas6 (also known as Csy4), which is the RNase responsible for processing of the precursor RNA transcript derived from the CRISPR locus. The Cas6–Cas7.1 interaction is mediated by a β-sheet formed by the contribution of a β-strand from Cas6 and the two β-strands that form the finger of Cas7.1 (Extended Data Fig. 5f). Cas6 also forms extensive interactions with the conserved stem-loop in the repeat-derived 3′ crRNA handle (Fig. 1, Extended Data Fig. 5d, e), with an arginine-rich α-helix (residues 110 to 128) docked in the major groove, positioning multiple basic residues within interaction distance of the negatively charged RNA backbone.

The interaction established between Cas6 and Cas7.1 forms a continuous surface on which TniQ is docked, forming the other protuberance of the crescent. The intrinsic flexibility of the complex resulted in lower local resolutions in this area of the maps, which we overcame using local alignments masking the area comprising TniQ, Cas6, Cas7.1 and the crRNA handle (Extended Data Fig. 7). The enhanced maps enabled de novo modelling and refinement of TniQ, for which no previous structure or homology model has been reported, to our knowledge (Fig. 2). Notably, TniQ binds to Cascade as a dimer with head-to-tail configuration (Fig. 2), a surprising result given the expectation that *E. coli* TnsD functions as a monomer during Tn*7* transposition[14].

TniQ binds to Cascade as a dimer

TniQ is composed of two domains: an N-terminal domain of approximately 100 residues formed by three short α-helices and a second, larger domain of approximately 300 residues with a signature sequence for the TniQ family. A DALI search[15] using the refined TniQ model as a probe yielded marked structural similarity of the N-terminal domain to proteins containing helix–turn–helix (HTH) domains. This domain is often involved in nucleic acid recognition; however, there are examples where it has been re-purposed for protein–protein interactions[16]. The remaining C-terminal TniQ domain is formed by ten α-helices of variable length and is predicted to contain two tandem zinc finger motifs, although this region was poorly defined in the maps (Fig. 2). Overall, the double domain composition of TniQ results in an elongated structure, bent at the junction of the HTH and the TniQ domain (Fig. 2). The HTH domain of one monomer engages the TniQ domain of the other monomer via interactions between α-helix 3 (H3) and α-helix 12 (H12), respectively, in a tight protein–protein interaction (Fig. 2c). This reciprocal interaction is complemented by multiple interactions established between the TniQ domains from both monomers (up to 45 non-covalent interactions as reported by PISA[17]).

Tethering of the TniQ dimer to Cascade is accomplished by specific interactions established with both Cas6 and Cas7.1 (Fig. 3). One monomer of TniQ interacts with Cas6 via its C-terminal TniQ domain, whereas the other TniQ monomer contacts Cas7.1 through its N-terminal HTH domain (Figs. 2b, 3). The loop connecting α-helices H7 and H8 of the TniQ domain of the first TniQ monomer is inserted in a hydrophobic cavity formed at the interface of two α-helices of Cas6 (Fig. 3b, d). The TniQ histidine residue 265 is involved in rearranging the hydrophobic loop connecting H7 and H8 (Fig. 3d), which is inserted in the hydrophobic pocket of Cas6 formed by residues L20, Y74, M78, Y83 and F84. The buried surface in the Cas6–TniQ.1 interaction interface has an area of 420 Å^2. The HTH domain of the other TniQ monomer interacts with Cas7.1 through a network of interactions established mainly by α-helix H2 and the linker connecting H2 and H3, burying a surface area of 595 Å^2 (Fig. 3c, e). Thus, the HTH domain and the TniQ domain exert dual roles to drive TniQ dimerization and dock onto Cascade. The aggregate buried surface area for the TniQ–Cascade interaction is 1,015 Å^2, significantly smaller than other Cascade–effector interactions such as with the nuclease Cas3, in which 2,433 Å^2 is buried[18]. This difference is not surprising given the flexibility observed for the TniQ dimer in its association with Cascade (Supplementary Video 1).

Structure of the DNA-bound TniQ–Cascade complex

To investigate the structural determinants of DNA recognition by the TniQ–Cascade complex, we determined the structure of the complex

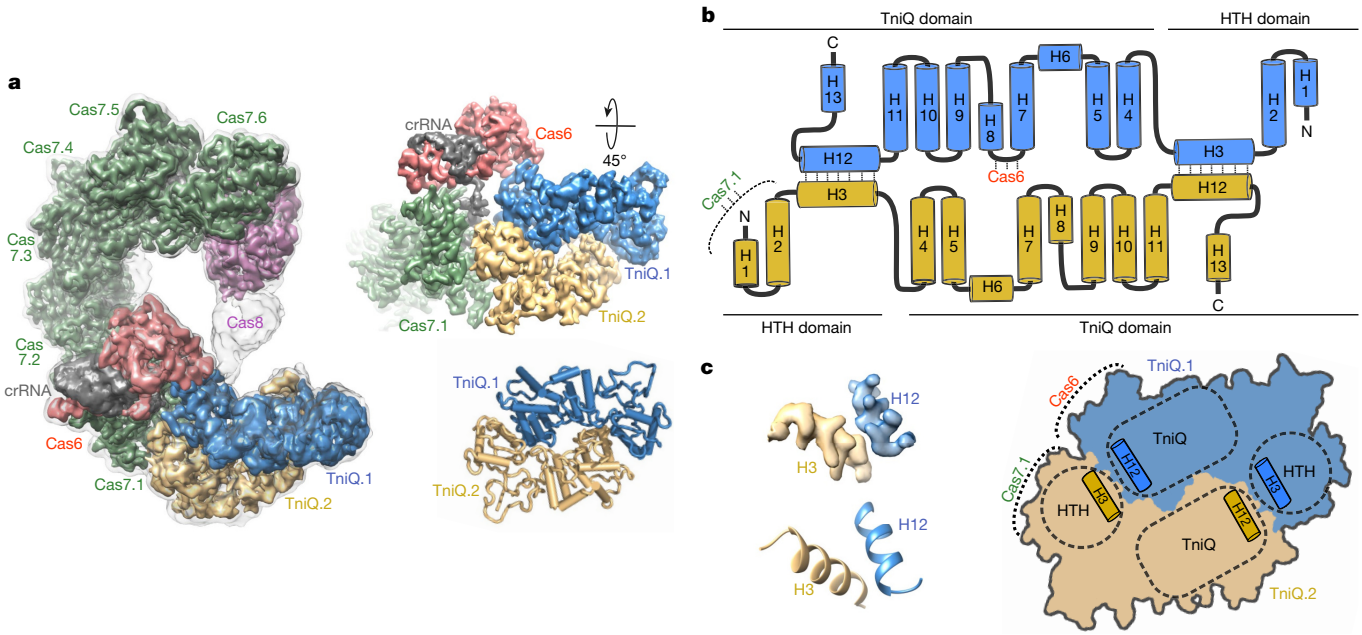


Fig. 2 | TniQ binds Cascade in a dimeric, head-to-tail configuration. a, Left, overall view of the TniQ–Cascade cryo-EM unsharpened map (grey) overlaid on the post-processed map segmented and coloured as in Fig. 1. Right, cryo-EM map (top) and refined model (bottom) of the TniQ dimer. The two monomers interact with each other in a head-to-tail configuration and are anchored to Cascade by Cas6 and Cas7.1. **b**, Secondary structure diagram of the TniQ dimer: thirteen α-helices are organized into an N-terminal HTH domain and a C-terminal TniQ domain. Dimer interactions between H3 and H12 are indicated, as are interaction sites with Cas6 and Cas7.1. **c**, Cryo-EM density for the H3–H12 interaction shows clear side-chain features (top), allowing accurate modelling of the interaction (bottom). **d**, Schematic of the dimer interaction, showing the important dimerization interface between the HTH and TniQ domain.

bound to a double-stranded DNA (dsDNA) substrate containing the 32-bp target sequence, 5′-CC-3′ PAM, and 20 bp of flanking dsDNA on both ends (Fig. 4, Extended Data Fig. 8). Density for 28 nucleotides of the target strand and 8 nucleotides of the non-target strand could be confidently assigned in the reconstructed maps (Fig. 4c). As with previous I-F Cascade structures, Cas8 recognizes the double-stranded PAM within the minor groove[11] (Extended Data Fig. 9), and an arginine residue (R246) establishes a stacking interaction with a guanine nucleotide on the target strand, which acts as a wedge to separate the double-stranded PAM from the neighbouring unwound DNA where base-pairing with the crRNA begins (Fig. 4c).

Twenty-two nucleotides of the target strand within the 32-bp target showed clear density, but surprisingly, the terminal nine nucleotides were not ordered. The target-strand base pairs with the spacer region of the crRNA in short, discontinuous helical segments, as observed previously for I-E and I-F DNA-bound Cascade complexes[11,12], with every sixth base flipped out of the heteroduplex by the insertion of a Cas7 finger (Extended Data Fig. 6b). The observed 22-bp heteroduplex is stabilized by the four Cas7 monomers proximal to the PAM (Cas7.6–Cas7.3), but even after local masked refinements, no density could be observed for any target strand nucleotides that would base-pair with the 3′ end of the crRNA spacer bound by Cas7.2 and Cas7.1. These two Cas7 monomers are proximal to Cas6 and in the region previously described to exhibit dynamics owing to the interaction of the Cas8 flexible domain with the inner face of the TniQ dimer. In addition, the disordered nucleotides also correspond to positions 25–28 of the target site where RNA–DNA mismatches

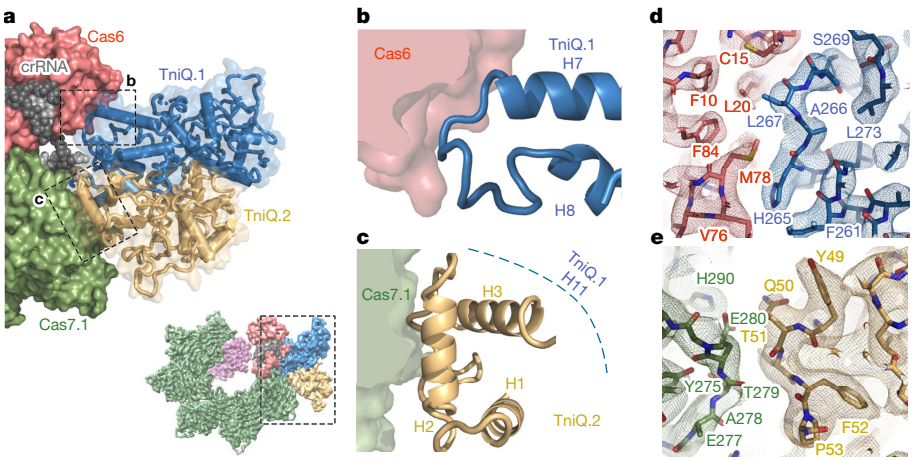


Fig. 3 | Cas6 and Cas7.1 form a binding platform for TniQ. a, Top, magnified area showing the interaction site of Cascade and the TniQ dimer. Cas6 and Cas7.1 are displayed as molecular Van der Waals surfaces, the crRNA is shown as grey spheres and the TniQ monomers are shown as ribbons. **b**, The loop connecting TniQ.1 α-helices H7 and H8 (blue) binds within a hydrophobic cavity of Cas6. **c**, Cas7.1 interacts with the HTH domain of the TniQ.2 monomer (yellow), mainly through H2 and the loop connecting H2 and H3. **d**, **e**, Experimental cryo-EM densities observed for the TniQ–Cas6 (**d**) and TniQ–Cas7.1 (**e**) interactions.

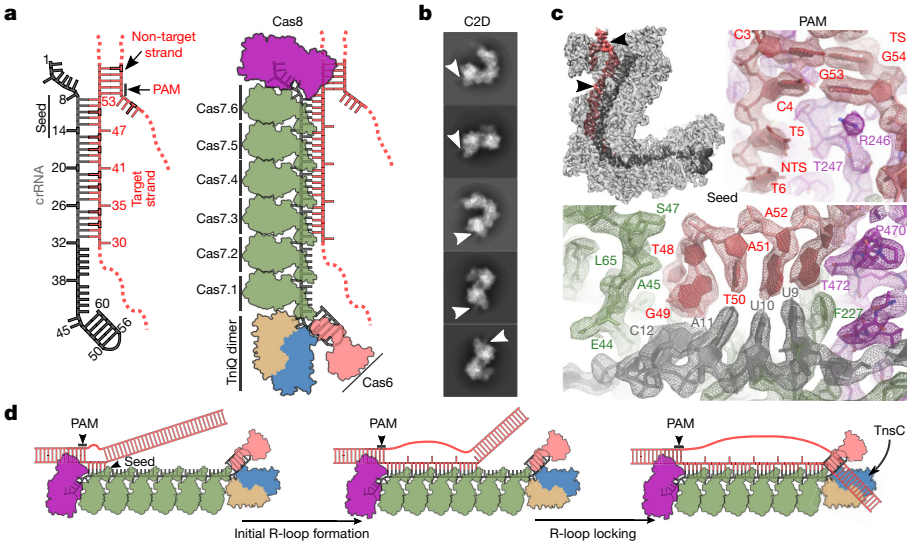


Fig. 4 | DNA-bound structure of the TniQ–Cascade complex. a, Schematic of crRNA and the portion of the dsDNA substrate that was experimentally observed within the electron density map for DNA-bound TniQ–Cascade. The target strand, non-target strand, PAM and seed regions are indicated (left); protein components are shown on the right. **b**, Selected cryo-EM reference-free two-dimensional class averages for DNA-bound TniQ–Cascade; density corresponding to dsDNA could be directly observed protruding from the Cas8 component in the two-dimensional class averages (white arrows). **c**, Cryo-EM map for DNA-bound TniQ–Cascade. The crRNA is shown in dark grey and the DNA is shown in red. Right bottom, detailed views of the PAM and seed-recognition regions of the map, with refined models represented as sticks within the electron density. Cas8 is shown in purple, Cas7 is shown in green, the crRNA is in grey and DNA is shown in red. NTS, non-target strand. **d**, The *V. cholerae* transposon encodes a TniQ–Cascade co-complex that uses the sequence content of the crRNA to bind complementary DNA target sites. We propose that the incomplete R-loop observed in our structure (middle) represents an intermediate state that may precede a downstream 'locking' step involving proofreading of the RNA–DNA complementarity. TniQ is positioned at the PAM-distal end of the DNA-bound Cascade complex, where it probably interacts with TnsC during downstream steps of RNA-guided DNA insertion.

are detrimental for RNA-guided DNA integration[7]. Thus, we propose that the partial R-loop structure that we observed could represent an intermediate conformation refractory to integration, and that further structural rearrangements may be critical for further stabilization of an open conformation, possibly driven by recruitment of the TnsC ATPase.

Here we present cryo-EM structures of a CRISPR–Cas effector complex bound to the transposition protein TniQ, with and without target DNA. These structures reveal the unexpected presence of TniQ as a dimer that forms bipartite interactions with Cas6 and Cas7.1 within the Cascade complex, forming a probable recruitment platform for downstream-acting transposition proteins[19] (Fig. 4d). Our structures further reveal a possible fidelity checkpoint, whereby formation of a complete R-loop requires conformational rearrangements that may depend on extensive RNA–DNA complementarity and/or downstream factor recruitment; this proofreading step could account for the highly specific RNA-guided DNA integration that we previously reported for the *V. cholerae* transposon[7]. In light of recent work demonstrating exaptation of type V-K CRISPR–Cas systems by similar Tn*7*-like transposons that also encode TniQ[20,21], it will be informative to determine whether tethering of TniQ to evolutionarily distinct crRNA effector complexes—Cascade or Cas12k—is a general theme of RNA-guided transposition.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1849-0.

1. Dy, R. L., Richter, C., Salmond, G. P. C. & Fineran, P. C. Remarkable mechanisms in microbes to resist phage infections. *Annu. Rev. Virol.* **1**, 307–331 (2014).
2. Hille, F. et al. The biology of CRISPR–Cas: backward and forward. *Cell* **172**, 1239–1259 (2018).
3. Doron, S. et al. Systematic discovery of antiphage defense systems in the microbial pangenome. *Science* **359**, eaar4120 (2018).
4. Sinkunas, T. et al. In vitro reconstitution of Cascade-mediated CRISPR immunity in *Streptococcus thermophilus*. *EMBO J.* **32**, 385–394 (2013).
5. Redding, S. et al. Surveillance and processing of foreign DNA by the *Escherichia coli* CRISPR–Cas system. *Cell* **163**, 854–865 (2015).
6. Peters, J. E., Makarova, K. S., Shmakov, S. & Koonin, E. V. Recruitment of CRISPR–Cas systems by Tn7-like transposons. *Proc. Natl Acad. Sci. USA* **114**, E7358–E7366 (2017).
7. Klompe, S. E., Vo, P. L. H., Halpin-Healy, T. S. & Sternberg, S. H. Transposon-encoded CRISPR–Cas systems direct RNA-guided DNA integration. *Nature* **571**, 219–225 (2019).
8. Peters, J. E. Tn7. *Microbiol. Spectr.* **2**, https://doi.org/10.1128/microbiolspec.MDNA3-0010-2014 (2014).
9. Jackson, R. N. et al. Crystal structure of the CRISPR RNA-guided surveillance complex from *Escherichia coli*. *Science* **345**, 1473–1479 (2014).
10. Chowdhury, S. et al. Structure reveals mechanisms of viral suppressors that intercept a CRISPR RNA-guided surveillance complex. *Cell* **169**, 47–57 (2017).
11. Guo, T. W. et al. Cryo-EM structures reveal mechanism and inhibition of DNA targeting by a CRISPR–Cas surveillance complex. *Cell* **171**, 414–426 (2017).
12. Mulepati, S., Héroux, A. & Bailey, S. Crystal structure of a CRISPR RNA-guided surveillance complex bound to a ssDNA target. *Science* **345**, 1479–1484 (2014).
13. Zivanov, J. et al. New tools for automated high-resolution cryo-EM structure determination in RELION-3. *eLife* **7**, 163 (2018).
14. Holder, J. W. & Craig, N. L. Architecture of the Tn7 posttransposition complex: an elaborate nucleoprotein structure. *J. Mol. Biol.* **401**, 167–181 (2010).
15. Holm, L. & Laakso, L. M. Dali server update. *Nucleic Acids Res.* **44** (W1), W351–W355 (2016).
16. Aravind, L., Anantharaman, V., Balaji, S., Babu, M. M. & Iyer, L. M. The many faces of the helix–turn–helix domain: transcription regulation and beyond. *FEMS Microbiol. Rev.* **29**, 231–262 (2005).
17. Krissinel, E. Stock-based detection of protein oligomeric states in jsPISA. *Nucleic Acids Res.* **43** (W1), W314–W319 (2015).
18. Xiao, Y., Luo, M., Dolan, A. E., Liao, M. & Ke, A. Structure basis for RNA-guided DNA degradation by Cascade and Cas3. *Science* **361**, eaat0839 (2018).
19. Choi, K. Y., Spencer, J. M. & Craig, N. L. The Tn7 transposition regulator TnsC interacts with the transposase subunit TnsB and target selector TnsD. *Proc. Natl Acad. Sci. USA* **111**, E2858–E2865 (2014).
20. Faure, G. et al. CRISPR–Cas in mobile genetic elements: counter-defence and beyond. *Nat. Rev. Microbiol.* **17**, 513–525 (2019).
21. Strecker, J. et al. RNA-guided DNA insertion with CRISPR-associated transposases. *Science* **365**, 48–53 (2019).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

No statistical methods were used to predetermine sample size. The experiments were not randomized. The investigators were not blinded to allocation during experiments and outcome assessment.

TniQ–Cascade purification

Protein components of TniQ–Cascade were expressed from a pET-derivative vector containing the native *V. cholerae tniQ-cas8-cas7-cas6* operon with an N-terminal His_{10}-MBP-TEV site fusion on TniQ. The crRNA was expressed separately from a pACYC-derivative vector containing a minimal repeat–spacer–repeat CRISPR array encoding a spacer from the endogenous *V. cholerae* CRISPR array. The TniQ–Cascade complex was overexpressed and purified as described previously[7], and was stored in Cascade storage buffer (20 mM Tris-Cl, pH 7.5, 200 mM NaCl, 1 mM DTT, 5% glycerol).

Sample preparation for electron microscopy

For negative staining, 3 µl of purified TniQ–Cascade ranging from 100 nM to 2 µM was incubated with plasma treated (H_2/O_2 gas mix, Gatan Solarus) CF400 carbon-coated grids (EMS) for 1 min. Excess solution was blotted and 3 µl of 0.75% uranyl formate was added for an additional minute. Excess stain was blotted away and grids were air-dried overnight. Grid screening for both negative staining and cryo conditions was performed on a Tecnai-F20 microscope (FEI) operated at 200 KeV and equipped with a Gatan K2-Summit direct detector. Microscope operation and data collection were carried out using the Leginon/Appion software. Initial negative staining grid screening allowed determination of a suitable concentration range for cryo conditions. Several grid geometries were tested in the 1–4 µM concentration range for cryo conditions using a Vitrobot Mark-II operated at 4 °C, 100% humidity, blot force 3, drain time 0, waiting time 15 s, and blotting times ranging from 3–5 s. The best ice distribution and particle density was obtained with 0.6/1 UltrAuFoil grids (Quantifoil).

Electron microscopy

A preliminary dataset of 300 images in cryo was collected with the Tecnai-F20 microscope using a pixel size of 1.22 Å/pixel with illumination conditions adjusted to 8 e^-/pixel/second with a frame window of 200 ms. Preprocessing and image processing were integrally done in Relion3[13] with ctf estimation integrated via a wrapper to Gctf[22]. An initial model computed using the SGD algorithm[23] implemented in Relion3 was used as initial reference for a refine three-dimensional job that generated a sub-nanometric reconstruction with approximately 10,000 selected particles. Clear secondary structure features in the two-dimensional averages and the three-dimensional reconstruction could be identified.

For the DNA-bound TniQ–Cascade complex containing DNA, we pre-incubated two complementary 74-nt oligonucleotides: (NTS: 5′-TTCATCAAGCCATTGGACCGCCTTACAGGACGCTTTGG CTTCATTGCTTTTCAGCTTCGCCTTGACGGCCAAAA-3′, TS: 5′-TTTTGG CCGTCAAGGCGAAGCTGAAAAGCAATGAAGCCAAAGCGTCCTGTAAGG CGGTCCAATGGCTTGATGAA-3′) for 5 min at 95 °C in hybridization buffer (20 mM Tris-Cl, pH 7.5, 100 mM KCl, 5 mM $MgCl_2$) to form dsDNA, which was subsequently aliquoted and flash-frozen. Complex formation was performed by incubating a 3× molar excess of dsDNA with TniQ–Cascade at 37 °C for 5 min before vitrification, which followed the conditions optimized for the apo complex (defined as TniQ–Cascade with crRNA but no DNA ligand).

High-resolution data for the apo complex were collected in a Tecnai-Polara-F30 microscope operated at 300 KeV equipped with a K3 direct detector (Gatan). A 30-µm C2 aperture was used with a pixel size of 0.95 Å/pixel and illumination conditions in microprobe mode adjusted to a fluence of 16 e^-/pixel/second. Four-second images with a frame width of 100 ms (1.77 e^-/Å^2/frame) were collected in counting mode.

For the DNA-bound complex, high-resolution data were collected in a Titan Krios microscope (FEI) equipped with an energy filter (20 eV slit width) and a K2 direct detector (Gatan) operated at 300 KeV. A 50-µm C2 aperture was used with a pixel size of 1.06 Å/pixel and illumination conditions adjusted in nanoprobe mode to a fluence of 8 e^-/pixel/second. Eight-second images with a frame width of 200 ms (1.42 e^-/Å^2/frame) were collected in counting mode.

Image processing

Motion correction was performed for every micrograph applying the algorithm described for Motioncor2[24] implemented in Relion3 with 5 by 5 patches for the K2 data and 7 by 5 patches for the K3 data. Parameters of the contrast transfer function for each motion-corrected micrograph were obtained using Gctf[22] integrated in Relion3. Initial particle picking of a subset of 200 images randomly chosen was performed with the Laplacian tool of the Auto-picking module of Relion3, using an estimated size for the complex of 200 Å. Then, 15,000 particles were extracted in a 300-pixel box size and binned 3 times for an initial two-dimensional classification job. Selected two-dimensional averages from this job were used as templates for Auto-picking of the full dataset. The full dataset of binned particles was subjected to a two-dimensional classification job to identify particles able to generate averages with clear secondary structure features. The selected subgroup of binned particles after the two-dimensional classification selection was refined against a three-dimensional volume obtained by SGD with the F20 data. This consensus volume was inspected to localize areas of heterogeneity that were clearly identified at both ends of the crescent shape characteristic of this complex. Both ends were then individually masked using soft masks of around 20 pixels that were subsequently used in classification jobs without alignments in Relion3. The T parameter used for this classification job was 6 and the total number of classes was 10. This strategy allowed us to identify two main population of particles which correspond to an open and closed state of the complex. Particles from both subgroups were separately re-extracted to obtain unbinned data sets for further refinement. New features implemented in Relion3, namely Bayesian polishing and ctf parameters refinement, allowed the extension of the resolution to 3.4, 3.5 and 2.9 Å for the two apo and the DNA-bound complexes, respectively. Post-processing was performed with a soft-mask of 5 pixels being the B-factor estimated automatically in Relion3 following standard practice. A final set of local refinements was performed with the masks used for classification. The locally aligned maps exhibit very good quality for the ends of the C-shape. These maps were used for de novo modelling and initial model refinement.

Model building and refinement

For the Cas7 and Cas6 monomers, the *E. coli* homologues (PDB accession code 4TVX) were initially docked with Chimera[25] and transformed to poly-alanine models. Substantial rearrangement of the finger region of Cas7 monomers, as well as other secondary structure elements of Cas6, were performed manually in COOT[26] before amino acid substitution of the poly-alanine model. Well-defined bulky side chains of aromatic residues allowed a confident assignment of the register. The crRNA was also well defined in the maps and was traced de novo with COOT. For Cas8 and TniQ in particular, no structural similarity was found in the published structures that was able to explain our densities. Locally refined maps using soft masks at both ends of the crescent-shaped complex rendered well-defined maps below 3.5 Å resolution. These maps were used for manual de novo tracing of a poly-alanine model in COOT that was subsequently mutated to the *V. cholerae* sequences. Bulky side chains for aromatic residues showed excellent density and were used as landmarks to adjust the register of the sequence.

For refinement, an initial step of real-space refinement against the cryo-EM maps was performed with the phenix.real_space refinement

tool of the Phenix package[27], with secondary structure restraints activated. A second step of reciprocal space refinement was performed in Refmac5[28], with secondary restraints calculated with Prosmart[29] and LibG[30]. Weight of the geometry term versus the experimental term was adjusted to avoid overfitting of the model into cryo-EM map, as previously reported[31]. Model validation was performed in Molprobity[32].

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Maps and models have been deposited in the Electron Microscopy Data Bank with accession codes EMD-20349, EMD-20350 and EMD-20351 and the Protein Data Bank with accession codes 6PIF, 6PIG and 6PIJ.

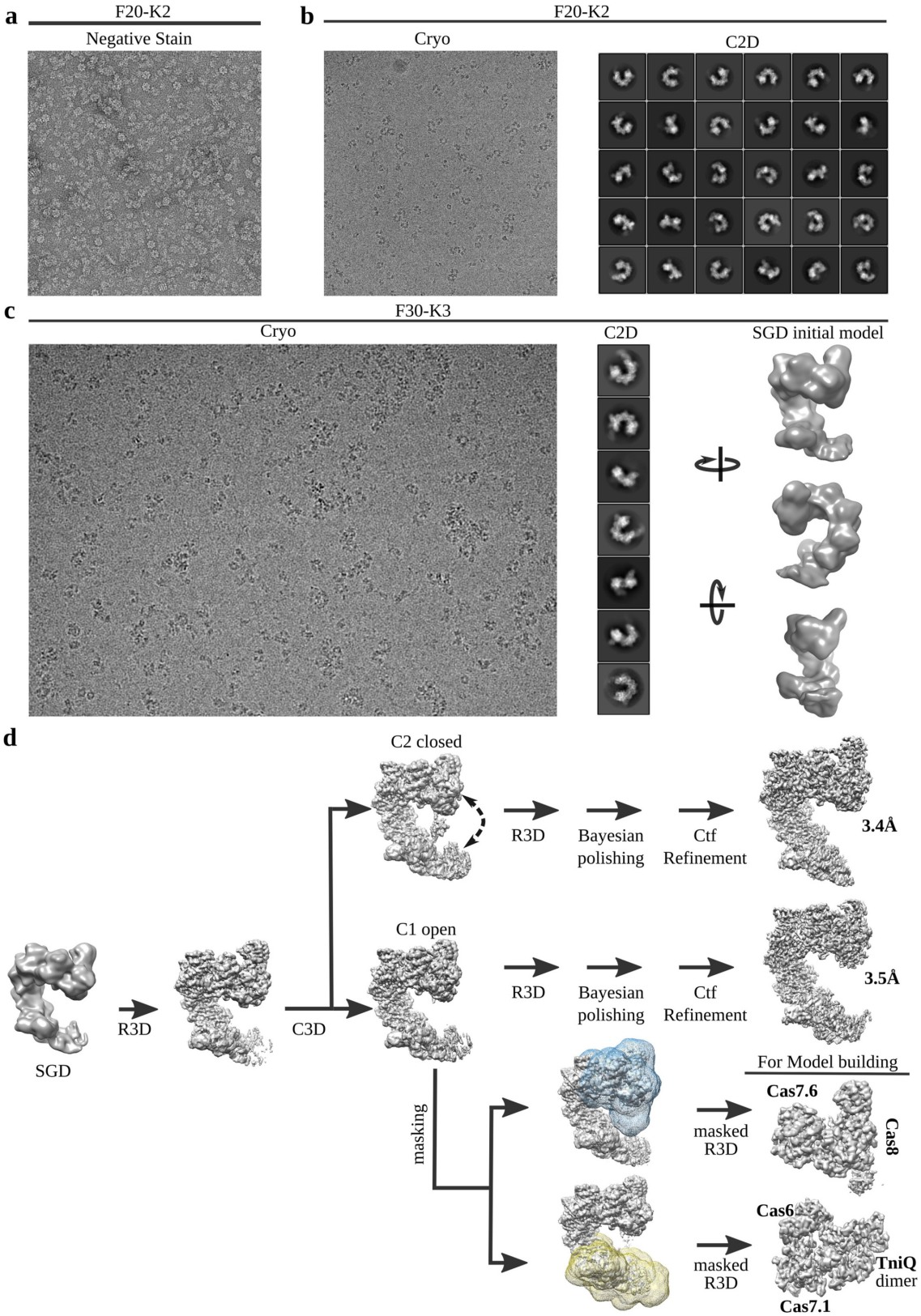
22. Zhang, K. Gctf: Real-time CTF determination and correction. *J. Struct. Biol.* **193**, 1–12 (2016).
23. Punjani, A., Rubinstein, J. L., Fleet, D. J. & Brubaker, M. A. cryoSPARC: algorithms for rapid unsupervised cryo-EM structure determination. *Nat. Methods* **14**, 290–296 (2017).
24. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
25. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
26. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
27. Afonine, P. V. et al. Real-space refinement in PHENIX for cryo-EM and crystallography. *Acta Crystallogr. D* **74**, 531–544 (2018).
28. Murshudov, G. N., Vagin, A. A. & Dodson, E. J. Refinement of macromolecular structures by the maximum-likelihood method. *Acta Crystallogr. D* **53**, 240–255 (1997).
29. Nicholls, R. A., Fischer, M., McNicholas, S. & Murshudov, G. N. Conformation-independent structural comparison of macromolecules with ProSMART. *Acta Crystallogr. D* **70**, 2487–2499 (2014).
30. Brown, A. et al. Tools for macromolecular model building and refinement into electron cryo-microscopy reconstructions. *Acta Crystallogr. D* **71**, 136–153 (2015).
31. Fernández, I. S., Bai, X.-C., Murshudov, G., Scheres, S. H. W. & Ramakrishnan, V. Initiation of translation by cricket paralysis virus IRES requires its translocation in the ribosome. *Cell* **157**, 823–831 (2014).
32. Williams, C. J. et al. MolProbity: more and better reference data for improved all-atom structure validation. *Protein Sci.* **27**, 293–315 (2018).
33. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).

Acknowledgements We thank B. Grassucci and Z. Zhang for technical assistance with cryo-EM data acquisition. Part of this work was performed at the Simons Electron Microscopy Center and National Resource for Automated Molecular Microscopy located at the New York Structural Biology Center, supported by grants from the Simons Foundation (SF349247), NYSTAR and the NIH National Institute of General Medical Sciences (GM103310).

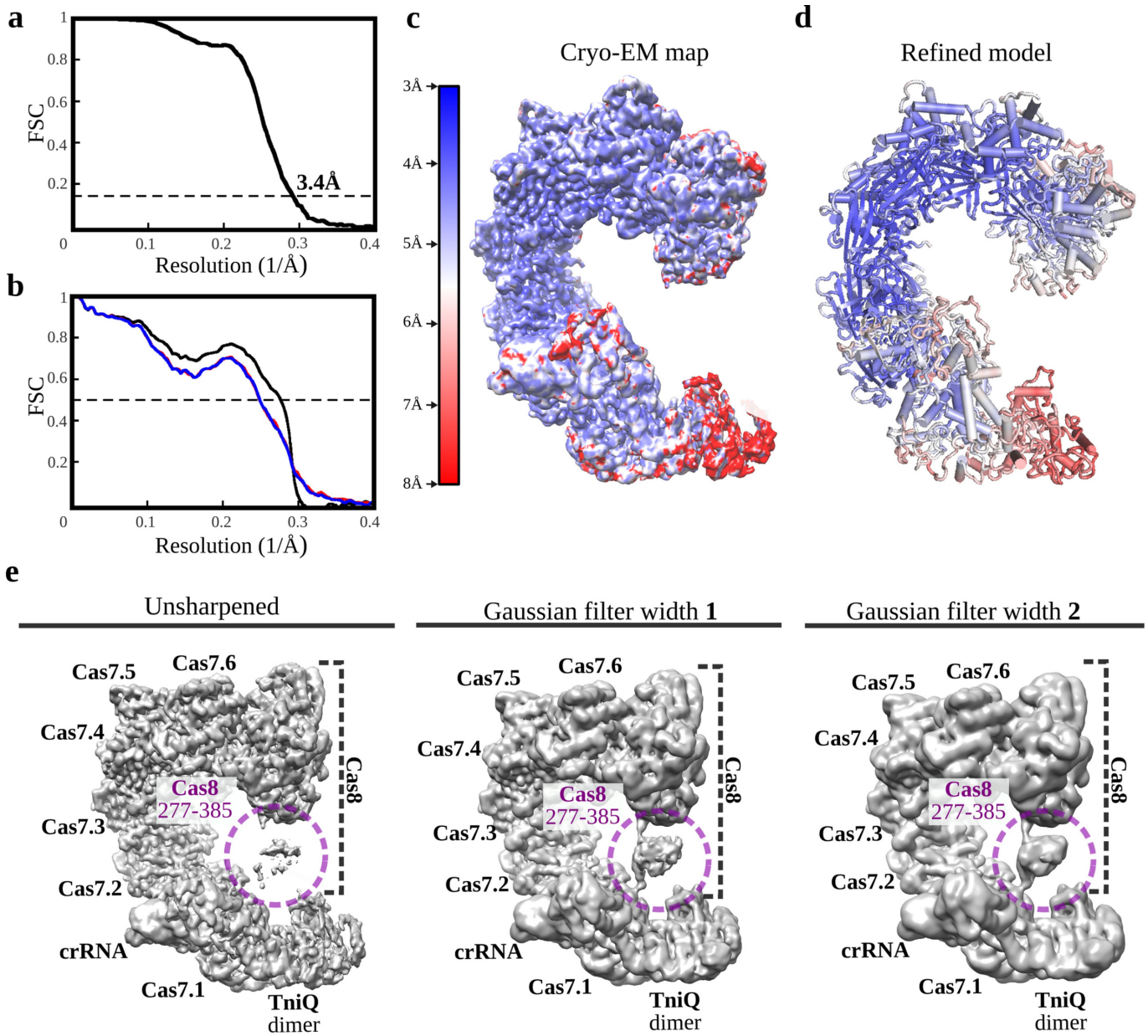
Author contributions All authors conceived and designed the project. T.S.H.-H. purified ribonucleoprotein complexes and assisted in cryo-EM data acquisition. I.S.F. collected EM data and solved the structures. I.S.F., S.H.S. and the other authors discussed the data and wrote the manuscript.

Competing interests Columbia University has filed a patent application related to this work. S.E.K. and S.H.S. are inventors on other patents and patent applications related to CRISPR–Cas systems and uses thereof. S.H.S. is a co-founder and scientific advisor to Dahlia Biosciences and an equity holder in Dahlia Biosciences and Caribou Biosciences.

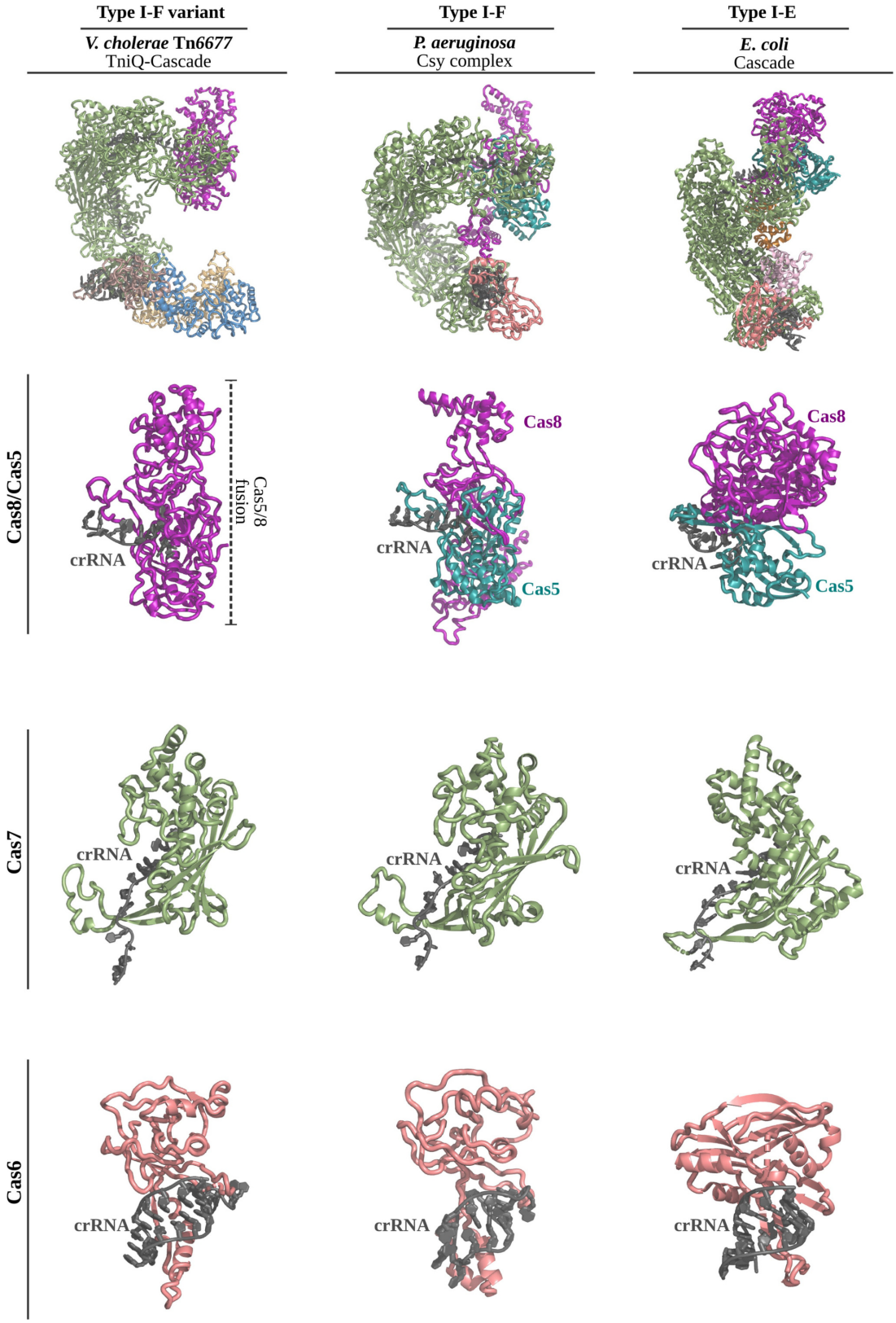
Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1849-0.
Correspondence and requests for materials should be addressed to S.H.S. or I.S.F.
Peer review information *Nature* thanks Scott Bailey, Ronald Chalmers, David Taylor and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.
Reprints and permissions information is available at http://www.nature.com/reprints.



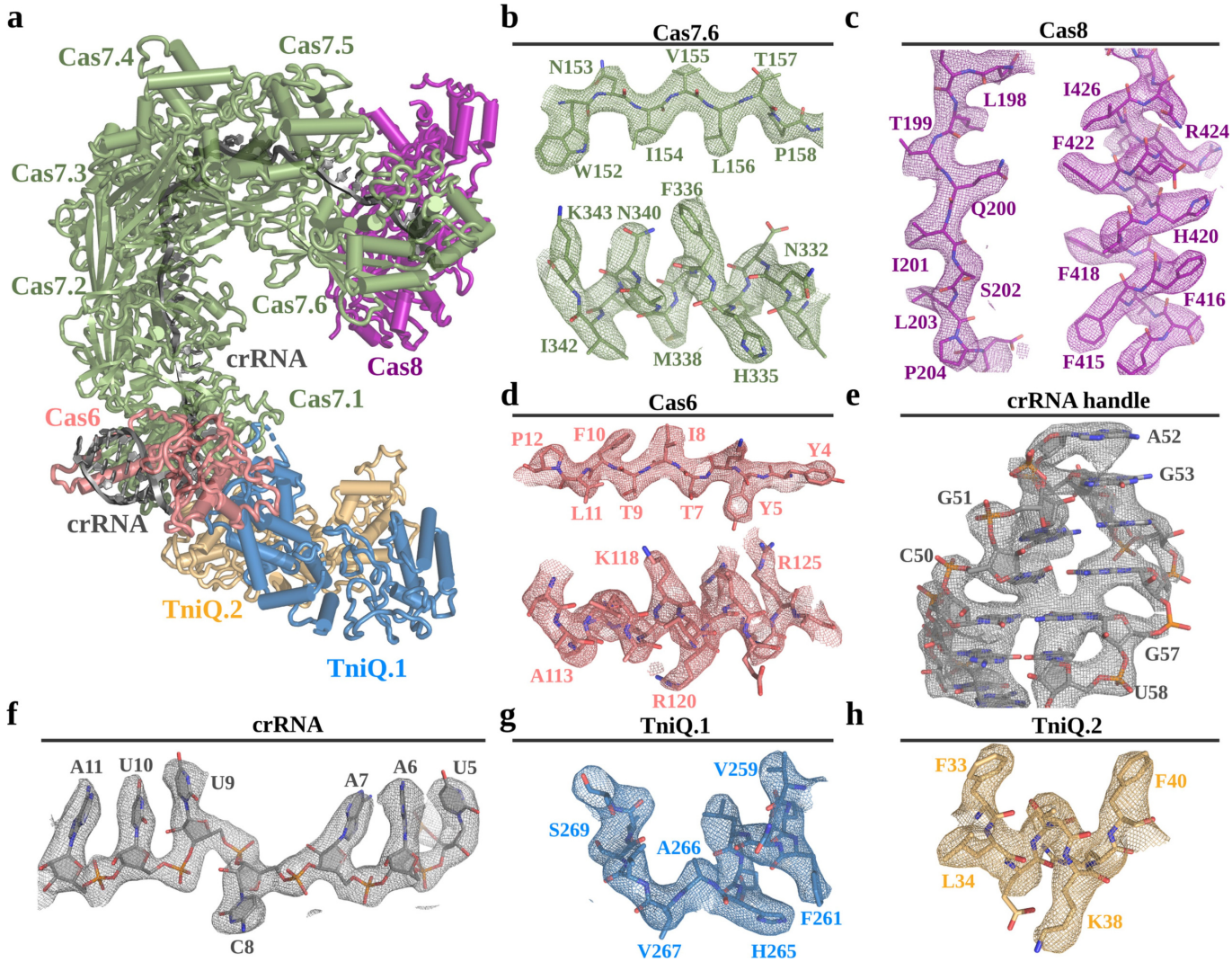
Extended Data Fig. 1 | Cryo-EM sample optimization and image processing workflow. a, Representative negatively stained micrograph for 500 nM TniQ–Cascade. **b**, Left, representative cryo-EM image for 2 µM TniQ–Cascade. A small dataset of 200 images was collected in a Tecnai-F20 microscope equipped with a Gatan K2 camera. Right, reference-free two-dimensional class averages for this initial cryo-EM dataset. **c**, Left, representative image from a large dataset collected in a Tecnai Polara microscope equipped with a Gatan K3 detector. Middle, detailed two-dimensional class averages were obtained that were used for initial model generation using the SGD algorithm[23] implemented in Relion3[13] (right). **d**, Image processing workflow used to identify the two main classes of the TniQ–cascade complex in open and closed conformations. Local refinements with soft masks were used to improve the quality of the map within the terminal protuberances of the complex. These maps were instrumental for de novo modelling and initial model refinement.



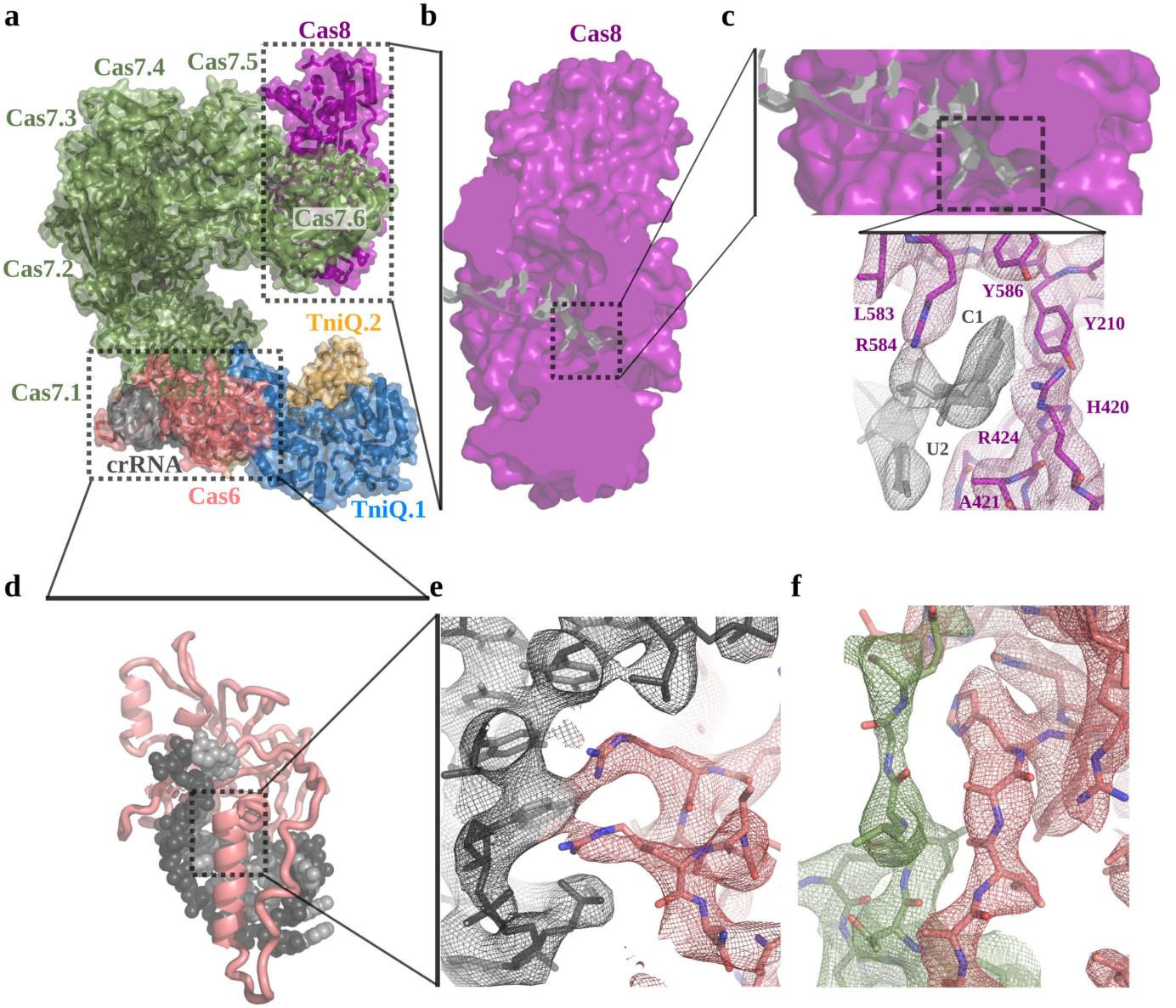
Extended Data Fig. 2 | Fourier shell correlation curves, local resolution, and unsharpened filter maps for the TniQ–Cascade complex in closed conformation. a, Gold-standard Fourier shell correlation (FSC) curve using half maps; the global resolution estimation is 3.4 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between the shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlap observed between the blue and red curves guarantees a non-overfitted model. **c**, Unsharpened map coloured according to local resolutions, as reported by RESMAP[33]. **d**, Final model coloured according to B-factors calculated by REFMAC[28]. **e**, A flexible Cas8 domain encompassing residues 277–385 contacts the TniQ dimer at the other side of the crescent shape. Applying a Gaussian filter of increasing width to the unsharpened map allows for a better visualization of this flexible region.



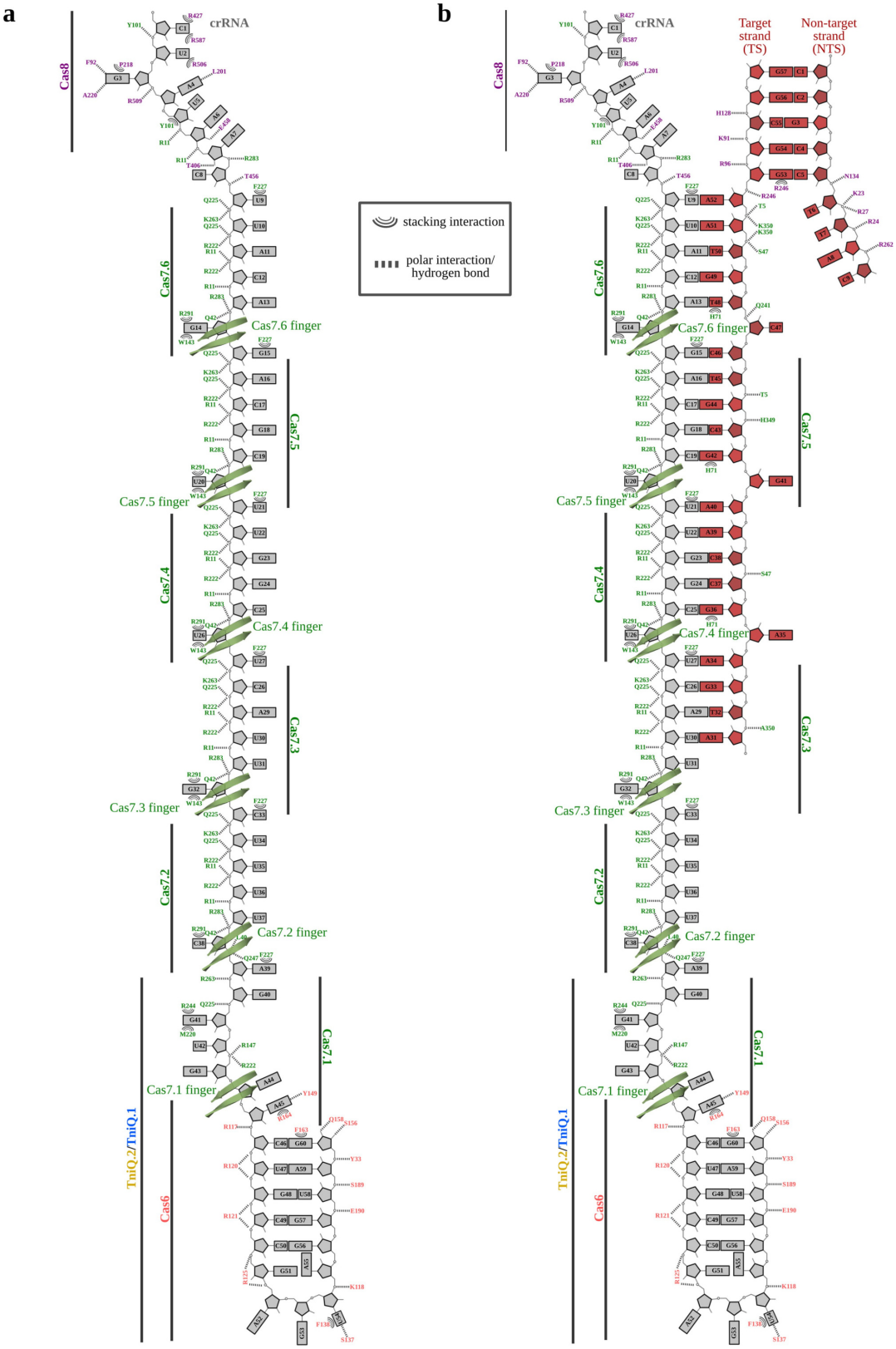
Extended Data Fig. 3 | Alignment of TniQ–Cascade with structurally similar Cascade complexes. The *V. cholerae* I-F variant TniQ–Cascade complex (left) was superposed with *Pseudomonas aeruginosa* I-F Cascade[11] (also known as Csy complex; middle, PDB ID: 6B45) and *E. coli* I-E Cascade[9] (right, PDB ID: 4TVX). Shown are alignments of the entire complex (top), the Cas8 and Cas5 subunits with the 5′ crRNA handle (second from top), the Cas7 subunit with a fragment of crRNA (second from bottom) and the Cas6 subunit with the 3′ crRNA handle (bottom).



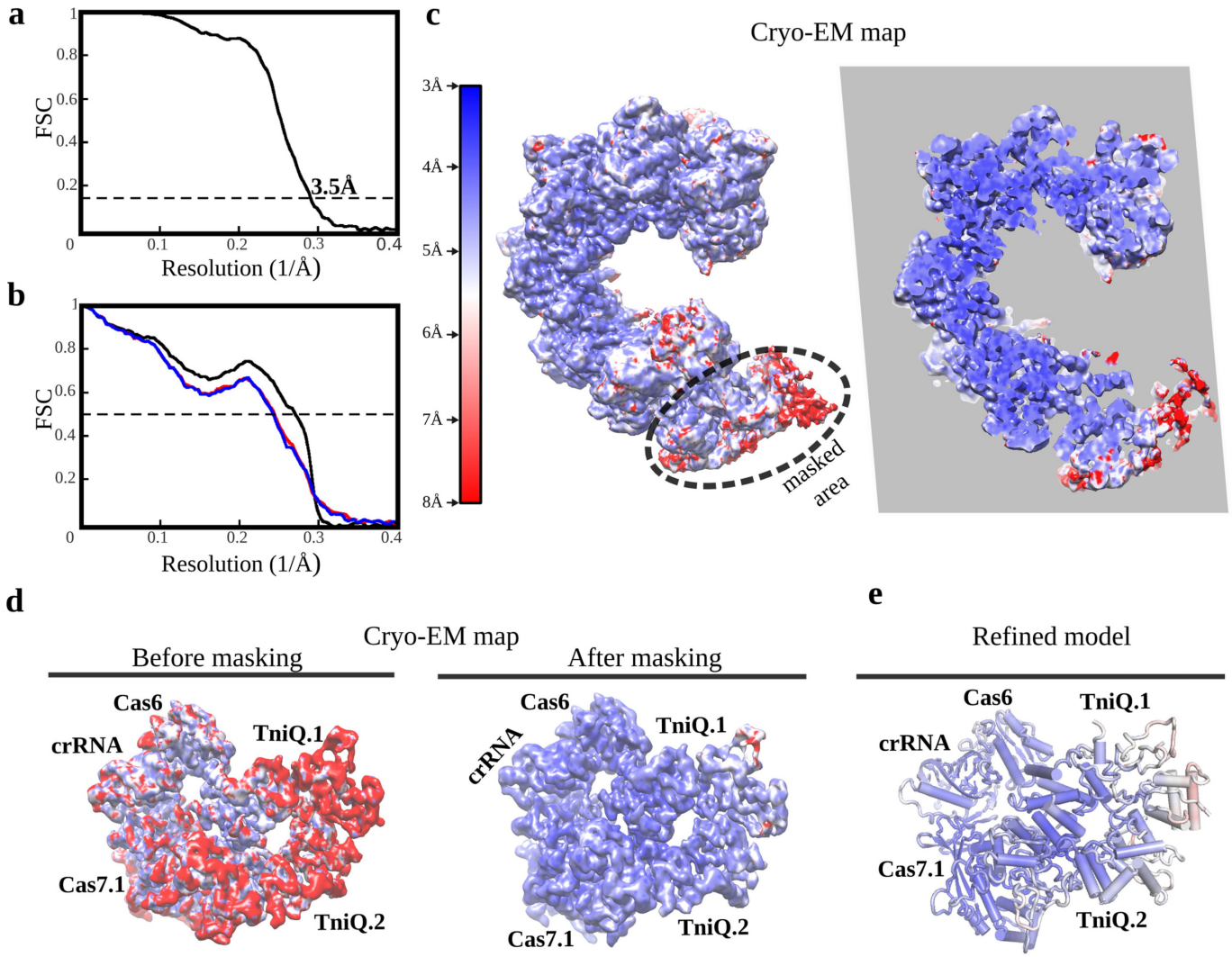
Extended Data Fig. 4 | Representative cryo-EM densities for all the components of the TniQ–Cascade complex in closed conformation. a, Final refined model of TniQ–Cascade, with Cas8 in purple, Cas7 monomers in green, Cas6 in salmon, the TniQ monomers in blue and yellow, and the crRNA in grey. **b**–**h**, Final refined model inserted in the final cryo-EM density for select regions of all the molecular components of the TniQ–Cascade complex. Residues are numbered.



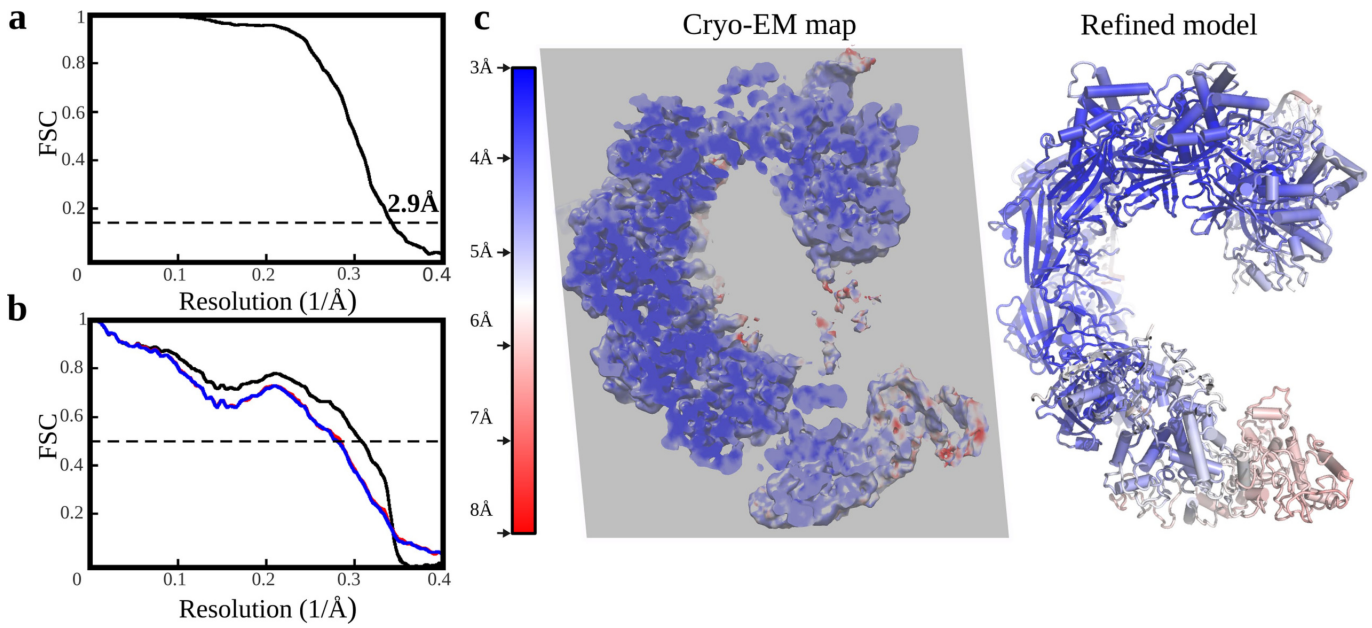
Extended Data Fig. 5 | Cas8 and Cas6 interaction with the crRNA. a, Refined model for the TniQ–Cascade shown as ribbons inserted in the semi-transparent Van der Waals surface, coloured as in Fig. 1. **b**, **c**, Magnified view of Cas8, which interacts with the 5′ end of the crRNA. The inset shows electron density for the highlighted region, where the base of nucleotide C1 is stabilized by stacking interactions with arginine residues R584 and R424. **d**, Cas6 interacts with the 3′ end of the crRNA 'handle' (nucleotides 45–60). **e**, An arginine-rich α-helix is deeply inserted within the major groove of the terminal stem–loop. This interaction is mediated by electrostatic interactions between basic residues of Cas6 and the negatively charged phosphate backbone of the crRNA. **f**, Cas6 (salmon) also interacts with Cas7.1 (green), establishing a β-sheet formed by β-strands contributed from both proteins.



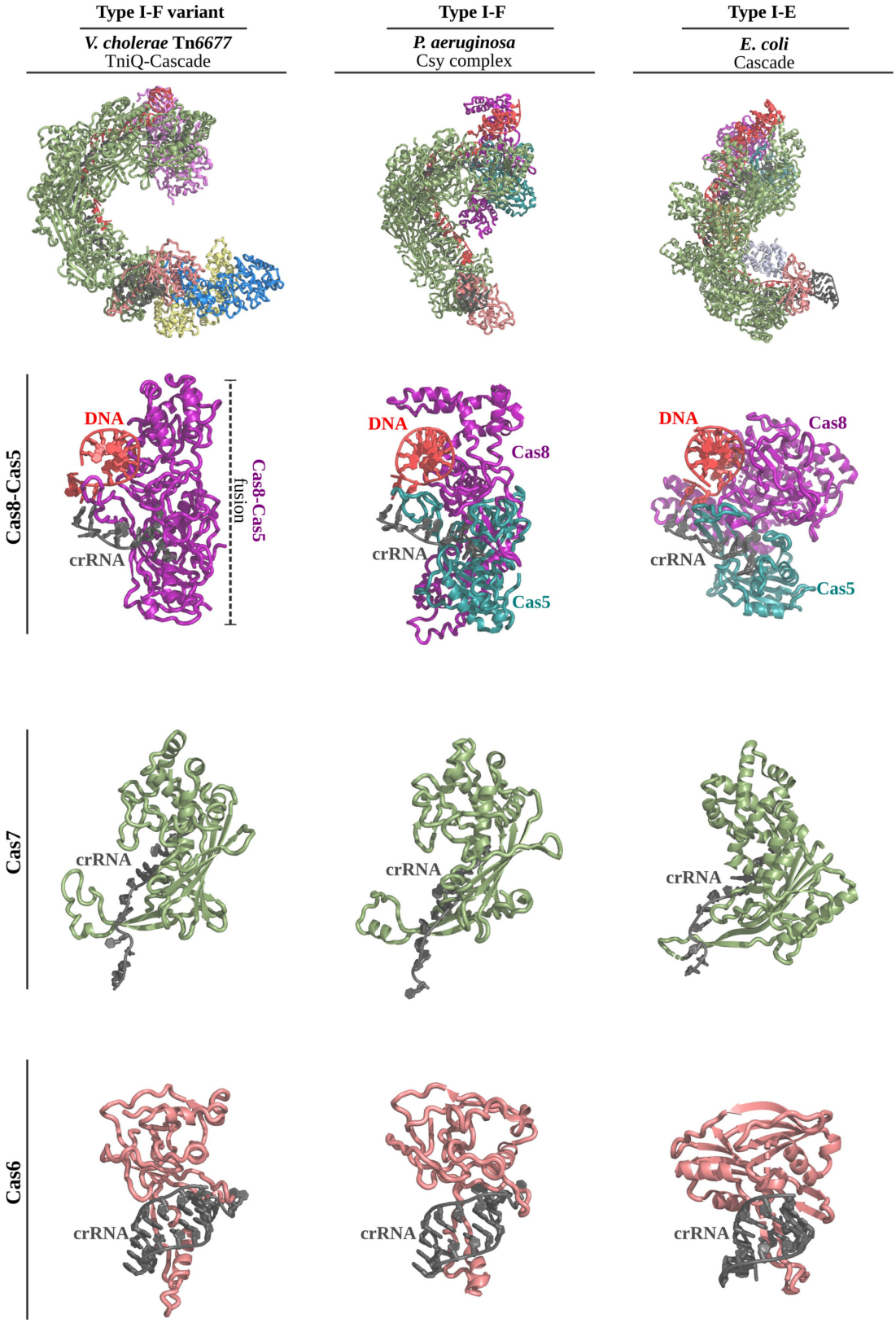
Extended Data Fig. 6 | Schematic representation of crRNA and target DNA recognition by TniQ–Cascade. a, TniQ–Cascade residues that interact with the crRNA are indicated. Approximate location for all protein components of the complex are also shown, as well as the position of each Cas7 finger. **b**, TniQ–Cascade residues that interact with crRNA and target DNA, shown as in **a**.



Extended Data Fig. 7 | FSC curves, local resolution, and local refined maps for the TniQ–Cascade complex in open conformation. a, Gold-standard FSC curve using half maps; the global resolution estimation is 3.5 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlapping between the blue and red curves guarantees a non-overfitted model. **c**, Unsharpened map coloured according to local resolutions, as reported by RESMAP[33]. Right, slice through the map shown on the left. **d**, Local refinements with soft masks improved the maps in flexible regions. Shown is the region of the map corresponding to the TniQ dimer. Unsharpened maps coloured according to the local resolution estimations are shown before (left) and after (right) masked refinements. **e**, Final model for the TniQ dimer region, coloured according to the local B-factors calculated by REFMAC[28].



Extended Data Fig. 8 | FSC curves, local resolution, and unsharpened filter maps for the DNA-bound TniQ–Cascade complex complex. a, Gold-standard FSC curve using half maps; the global resolution estimation is 2.9 Å by the FSC 0.143 criterion. **b**, Cross-validation model-versus-map FSC. Blue curve, FSC between the shacked model refined against half map 1; red curve, FSC against half map 2, not included in the refinement; black curve, FSC between final model against the final map. The overlap observed between the blue and red curves guarantees a non-overfitted model. **c**, Left, unsharpened map coloured according to local resolutions, as reported by RESMAP[33]. dsDNA is visible at the top right projecting outside of the complex. Right, final model coloured according to B-factors calculated by REFMAC[28].



Extended Data Fig. 9 | Alignment of DNA-bound TniQ–Cascade with structurally similar Cascade complexes. The DNA-bound structure of *V. cholerae* I-F variant TniQ–Cascade complex (left) was superposed with DNA-bound structures of *P. aeruginosa* I-F Cascade[11] (also known as Csy complex; middle, PDB ID: 6B44) and *E. coli* I-E Cascade[9] (right, PDB ID: 5H9F). Shown are alignments of the entire complex (top), the Cas8 and Cas5 subunits with the 5′ crRNA handle and double-stranded PAM DNA (middle top), the Cas7 subunit with a fragment of crRNA (middle bottom), and the Cas6 subunit with the 3′ crRNA handle (bottom).

nature research

Corresponding author(s): Israel S. Fernandez

Last updated by author(s): Aug 13, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	For cryoEM data collection we have used the Leginon/Appion package.
Data analysis	Ctf estimation was performed with Gctf v1.06 and Image processin with Relion v3.0-beta. Figures were generated with Chimera and PyMol.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

PDBs and cryoEM maps are deposited in the PDB and EMDB public databases with accession codes: 6PIF, 6PIG and 6PIJ.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	*Describe how sample size was determined, detailing any statistical methods used to predetermine sample size OR if no sample-size calculation was performed, describe how sample sizes were chosen and provide a rationale for why these sample sizes are sufficient.*
Data exclusions	*Describe any data exclusions. If no data were excluded from the analyses, state so OR if data were excluded, describe the exclusions and the rationale behind them, indicating whether exclusion criteria were pre-established.*
Replication	*Describe the measures taken to verify the reproducibility of the experimental findings. If all attempts at replication were successful, confirm this OR if there are any findings that were not replicated or cannot be reproduced, note this and describe why.*
Randomization	*Describe how samples/organisms/participants were allocated into experimental groups. If allocation was not random, describe how covariates were controlled OR if this is not relevant to your study, explain why.*
Blinding	*Describe whether the investigators were blinded to group allocation during data collection and/or analysis. If blinding was not possible, describe why OR explain why blinding was not relevant to your study.*

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Article

Structure of the human metapneumovirus polymerase phosphoprotein complex

https://doi.org/10.1038/s41586-019-1759-1

Received: 1 July 2019

Accepted: 30 October 2019

Published online: 7 November 2019

Junhua Pan[1,2,8*], Xinlei Qian[3,4,5,8], Simon Lattmann[4], Abbas El Sahili[3,4], Tiong Han Yeo[3], Huan Jia[3,4], Tessa Cressey[6], Barbara Ludeke[6], Sarah Noton[6], Marian Kalocsay[7], Rachel Fearns[6,9*] & Julien Lescar[3,4,5,9*]

Respiratory syncytial virus (RSV) and human metapneumovirus (HMPV) cause severe respiratory diseases in infants and elderly adults[1]. No vaccine or effective antiviral therapy currently exists to control RSV or HMPV infections. During viral genome replication and transcription, the tetrameric phosphoprotein P serves as a crucial adaptor between the ribonucleoprotein template and the L protein, which has RNA-dependent RNA polymerase (RdRp), GDP polyribonucleotidyltransferase and cap-specific methyltransferase activities[2,3]. How P interacts with L and mediates the association with the free form of N and with the ribonucleoprotein is not clear for HMPV or other major human pathogens, including the viruses that cause measles, Ebola and rabies. Here we report a cryo-electron microscopy reconstruction that shows the ring-shaped structure of the polymerase and capping domains of HMPV-L bound to a tetramer of P. The connector and methyltransferase domains of L are mobile with respect to the core. The putative priming loop that is important for the initiation of RNA synthesis is fully retracted, which leaves space in the active-site cavity for RNA elongation. P interacts extensively with the N-terminal region of L, burying more than 4,016 Å^2 of the molecular surface area in the interface. Two of the four helices that form the coiled-coil tetramerization domain of P, and long C-terminal extensions projecting from these two helices, wrap around the L protein in a manner similar to tentacles. The structural versatility of the four P protomers—which are largely disordered in their free state—demonstrates an example of a 'folding-upon-partner-binding' mechanism for carrying out P adaptor functions. The structure shows that P has the potential to modulate multiple functions of L and these results should accelerate the design of specific antiviral drugs.

The 13-kilobase nonsegmented negative-strand RNA genome of HMPV encodes nine proteins[1]. Following virus entry, nucleoprotein-RNA is released into the cytoplasm of the host cell and serves as a template for the viral polymerase to produce sub-genomic capped and polyadenylated viral mRNAs, and for genome replication through an antigenome intermediate. A cap 1 structure is formed at the 5′ end of viral mRNAs by the capping and methyltransferase (MTase) domains of L[2,3]. Structures are available for segmented negative-strand virus polymerases such as influenza[4], which acquire a capped primer by 'cap snatching'. For polymerases of the *Mononegavirales* order, which includes several major human pathogens, the structure of the rhabdovirus vesicular stomatitis virus (VSV) L protein was determined first[3] and a crystal structure of the MTase domain of HMPV-L was reported[5].

Co-expressing the L and P proteins (Fig. 1a, b and Extended Data Fig. 1a) of HMPV in Sf9 insect cells yielded a stoichiometric preparation that was suitable for functional and structural studies (Fig. 1c–e). A primer elongation assay demonstrated that wild-type (WT) L_{WT}:P complexes had polymerase activity (Fig. 1c). L_{WT}:P complexes were also tested for de novo RNA synthesis activity using a template that consisted of a sequence from the 3′ end of the HMPV genome (le14). The major band that was detected in reactions, which contained L_{WT}:P, all four NTPs and radiolabelled UTP as a tracer, was 12 nucleotides (nt) in length, consistent with position 3C being the dominant initiation site (Fig. 1d). This product was not detectable in a reaction mix that contained only radiolabelled UTP and no other NTPs, or in a reaction that contained a chain terminator, confirming that it was the product of de novo initiation at the promoter. L(D745A):P, in which Asp745 from the conserved 'GDNQ' motif C was mutated to Ala, was devoid of polymerase activity, ruling out the possibility that a contaminating polymerase was responsible for the observed activity. Thus, recombinant L:P is active for RNA polymerization (Fig. 1c, d and Extended Data Fig. 2).

We determined the structure of the L:P complex to a resolution of 3.7 Å using cryo-electron microscopy (cryo-EM) (Fig. 1e, Extended

[1]Division of Molecular Medicine, Boston Children's Hospital, Boston, MA, USA. [2]Department of Biological Chemistry and Molecular Pharmacology, Harvard Medical School, Boston, MA, USA. [3]School of Biological Sciences, Nanyang Technological University, Singapore, Singapore. [4]NTU Institute of Structural Biology, Nanyang Technological University, Singapore, Singapore. [5]Antimicrobial Resistance Interdisciplinary Research Group, Singapore-MIT Alliance for Research and Technology, Singapore, Singapore. [6]Boston University School of Medicine, National Emerging Infectious Diseases Laboratories, Boston, MA, USA. [7]Laboratory of Systems Pharmacology, Department of Systems Biology, Harvard Medical School, Boston, MA, USA. [8]These authors contributed equally: Junhua Pan, Xinlei Qian. [9]These authors jointly supervised this work: Rachel Fearns, Julien Lescar. *e-mail: pan@crystal.harvard.edu; rfearns@bu.edu; julien@ntu.edu.sg

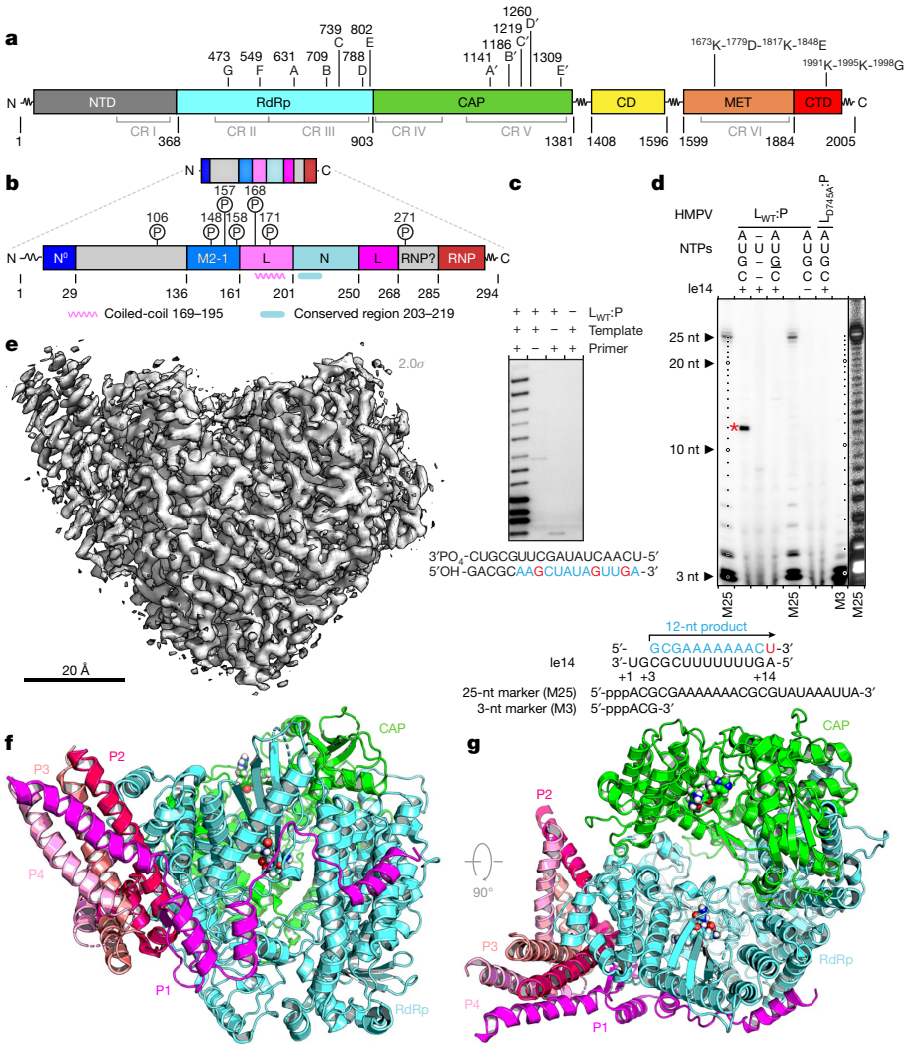


Fig. 1 | Overall structure of the HMPV-L:P complex. a, **b**, Domain organization of L (**a**) and P (**b**) outlining the conserved regions and motifs. Regions of HMPV-P that are predicted to interact with N^0, M2-1, L and RNP, and phosphorylation sites are indicated[11]. **c**, Primer elongation assay. Sequences of the 18-nt RNA template and for the 5-nt primer are shown with nascent RNA in blue and radiolabel incorporation sites in red. This experiment was performed a total of four times with two different buffer conditions. **d**, RdRp activity assay using the le14 RNA template. Sequences for the 5′-triphosphorylated 25-nt and 3-nt markers and the 12-nt product are indicated. The 3′dGTP chain terminator is labelled G. Radiolabelled UMP is shown in red. Data are representative of three independent experiments. For gel source data, see Supplementary Fig. 1. **e**, Overview of the HMPV-L:P cryo-EM three-dimensional reconstruction. **f**, Overview of the L:P atomic model. RdRp, cyan; capping domain, green; phosphoprotein tetramer subunits: P1, magenta; P2, hot pink; P3, salmon; P4, pink. Atoms of the GDNQ motif in the RdRp and the 'HR' motif in the capping domains are shown as coloured spheres. **g**, Rotated view of the L:P atomic model.

Data Fig. 1 and Supplementary Video 1). The final three-dimensional reconstruction (Extended Data Fig. 3) enabled us to build an atomic model for residues 8–1380 of L and for the bound P tetramer (Fig. 1f, g and Extended Data Figs. 3–5). Residues 1381 to the C terminus are not visible, suggesting flexibility of the peripheral appendage[5] comprising CR-VI+. A comparison of HMPV-L with VSV-L[3] returns a root mean square deviation (r.m.s.d.) of 2.8 Å, illustrating the preservation of the core L structure during the evolution of *Mononegavirales* (Supplementary Fig. 2). This similarity extends to the N-terminal domain (NTD) of VSV-L, which is essential for viral transcription[6] (Extended Data Fig. 6). The RdRp folds into the canonical fingers–palm–thumb subdomains observed in many RNA polymerase structures[3] (Fig. 2a). Nucleoside inhibitors have been developed that target the RdRp activity of RSV and HMPV[7]. The 5′ triphosphate metabolites of ALS-8176/ALS-8112 inhibit HMPV and RSV in nonhuman primates with similar potency and resistant RSV mutants have been sequenced[7]. We mapped the locations of RSV-L resistance mutations[7] to helix α29 of the present HMPV-L structure (Extended Data Fig. 7).

As in VSV-L, the capping domain of HMPV-L (Fig. 2b), containing the HR catalytic motif[2], has a large interface with the RdRp, accounting for the rigidity of the 'doughnut'. Arg1264 is exposed to the solvent, owing to the mobility of the C-terminal appendage. A feature shared by RdRps, which do not need a polynucleotide primer to initiate polymerization, is a priming loop (Fig. 2c, d) that supports the initiating rNTP[8]. In the influenza polymerase[4,9], a proline at the tip of a β-hairpin of PB1 has a key role in replication initiation. In rhabdoviruses (VSV and rabies virus), the corresponding loop is in the capping domain and contains residues that are important both for terminal de novo initiation and for mRNA capping[10], with a Trp residue in the capping domain that is essential for de novo initiation[10]—indicating it is the priming residue. In sequence alignments, a Tyr residue lies at a similar position in HMPV-L (Tyr1201) and RSV-L (Tyr1276), which could provide base-stacking interactions that are required for priming (Supplementary Fig. 2). We tested this possibility using the RSV polymerase, which initiates at positions 1 and 3 on its promoter. A Tyr1276Ala substitution did not hamper RSV initiation from either site, but the substitution of either Pro1261 or

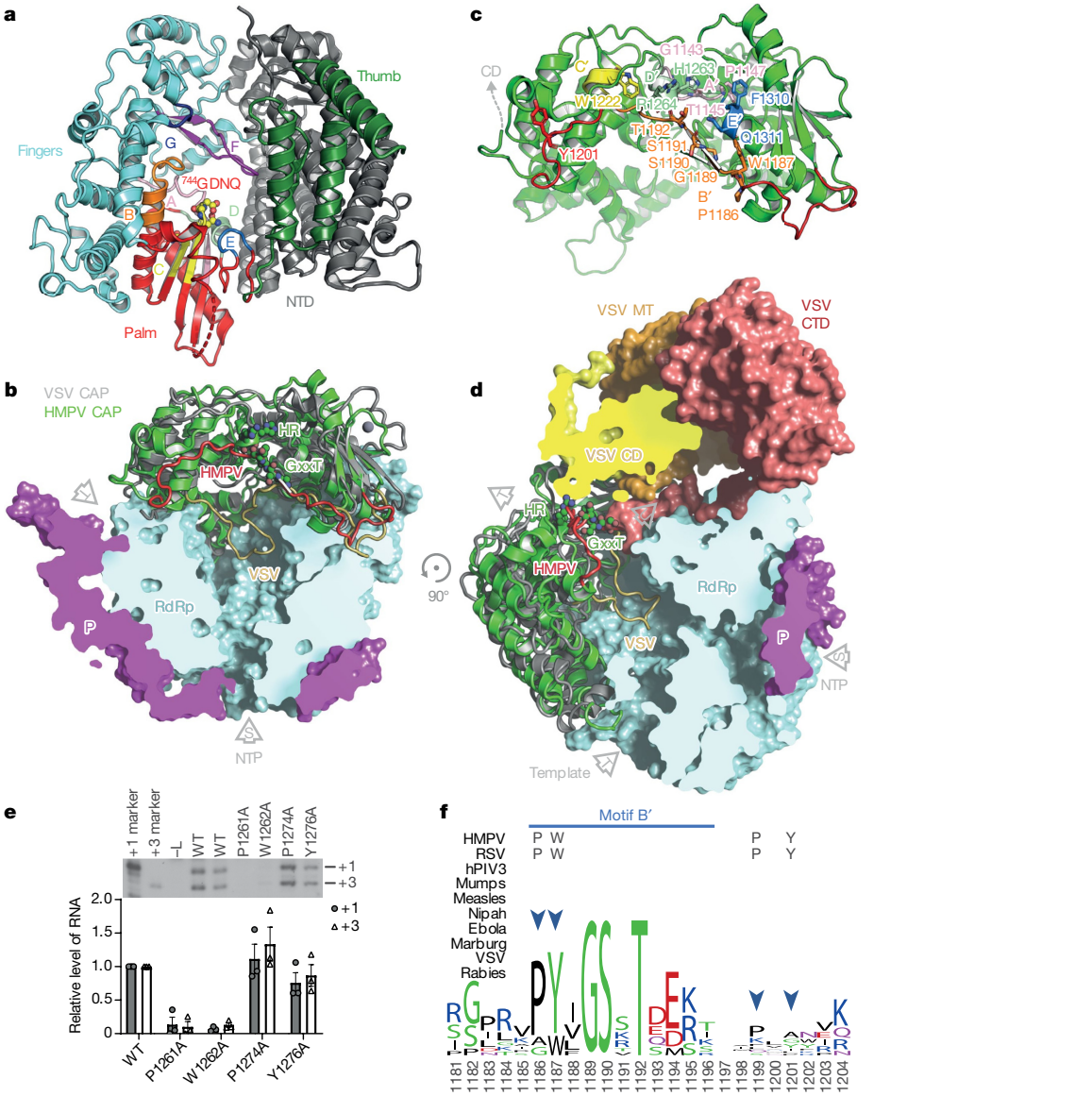


Fig. 2 | Structure of the HMPV-L protein. a, The RdRp viewed in its front orientation. Fingers subdomain, blue; palm, red; thumb, green; NTD, grey. RdRp motifs A–G are shown. **b**, The capping domain of L (green ribbons) of HMPV-L. A superposition of HMPV-L with VSV-L[3] (grey ribbons) highlights the difference between conformations of the putative priming loops. The putative priming loop of HMPV-L (including Thr1192 of the GXXT motif) is shown in red and of VSV-L (1157–1173) is shown in gold. P, magenta. **c**, RNA capping motifs[2] A′–E′. **d**, Overall view of HMPV-L with VSV-L superimposed, highlighting the mobile appendage (displayed in the VSV-L orientation, but disordered in the present structure) and rNTP (S) entry, template (T) and nascent (U) RNA tunnels. **e**, Effect of substitutions in aromatic and proline residues in the putative priming loop of RSV-L on RNA production from the +1 (circles) and +3 (triangles) sites. Data are mean ± s.e.m. of three independent experiments. **f**, Sequence conservation in the putative priming loop of L. The four residues that were mutated (**e**) are indicated by arrows. A proline residue (P1186 in HMPV-L) that is important for an early stage of RNA synthesis is conserved in *Pneumoviridae*, *Paramyxoviridae* and *Filoviridae* but not in *Rhabdoviridae*.

Trp1262 with Ala—corresponding to Pro1186 and Trp1187 in HMPV-L in motif B′ of the capping domain (Fig. 2c)—inhibited an early stage of RNA synthesis within the promoter region (Fig. 2e, f), suggesting that either this Pro or Trp residue might be responsible for priming de novo RNA synthesis. However, we cannot exclude the possibility that the N-terminal region of the loop has a structural function that is not directly related to de novo initiation in HMPV.

In VSV-L, the priming loop largely occludes the central RNA-binding cavity (Fig. 2d), precluding the formation of a long double-stranded RNA duplex and indicating a pre-initiation conformation[3]. The putative priming loop in our structure of HMPV-L retracts completely into a cavity in the capping domain, leaving ample space for several base pairs of a primer–template double-stranded RNA intermediate (Fig. 2b, d). Modelling an elongation complex in the HMPV-L tunnel shows that an RNA duplex can indeed be accommodated in the central cavity (Fig. 3a–d).

In the pre-initiation state of VSV-L, the connector domain (CD) and MTase domain have a defined orientation with respect to the core, being locked in place by the addition of peptide 35–106 of the VSV-P protein[3]. In HMPV-L, one linker region is located approximately at residues 1381–1407 just upstream of the CD and another is found at residues 1596–1598 before the CR-VI+ domain (Fig. 1a); these linker regions probably account for the flexibility of CD and CR-VI+. Mobility of the C-terminal domain (CTD), which has nucleoside 5′-triphosphatase, nucleoside-2′-*O*- and guanine-N[7]-MTase activities[5], might be necessary to approach the 5′ end of nascent mRNA for cap addition. Thus, it is possible that the N-terminal residues of HMPV-P (which are disordered here) are a latch that keeps the appendage in a fixed orientation for initiation and releases it during RNA elongation.

Clear density corresponding to a coiled-coil of four helices that constitute the P oligomerization region[11] enabled the tracing of residues 169–266 and 171–236 for the two proximal P subunits P1 and P2 in close contact with L, and of residues 168–219 and 169–231 for the two distal P subunits P3 and P4 (Figs. 1f, g, 4a). Much of P appears to fold upon binding to L, as shown by the presence of several extended arms that are typical of large molecular assemblies such as ribosomes or viral particles (Figs. 1, 4b and Supplementary Video 2). A large molecular surface area (Extended Data Fig. 4) becomes buried in the interaction, which mostly involves residues from P1 and P2, whereas P3 and P4 provide few interactions (Fig. 4a). Previous structural studies mapped the tetramerization and L-binding regions of P to the four parallel α-helices (spanning residues 171–193 of P) that could be visualized using X-ray crystallography[11]. By contrast, the N-terminal M2-1 binding regions of P[12], and the C-terminal regions of P that bind to the ribonucleoprotein (RNP) were found to be intrinsically disordered in the absence of their respective binding partners[11]. Although the tetramerization region of P appears as a rigid unit that is largely unaltered after binding to L, a disorder-to-order transition in the C-terminal regions of the two subunits proximal to L takes place upon binding: residues 194–213 of P1 fold into a β-hairpin, such that a 3-stranded antiparallel β-sheet is formed with residues 383–391 of the fingers subdomain of L (Fig. 4c). This β-hairpin is followed by a long tentacle-like structure that is composed of three consecutive α-helices that encircle the rear aperture, through which rNTPs gain access to the active-site chamber (Fig. 3a). Residues at the C terminus of the adjacent coiled-coil-forming helix (from P2) project in the opposite direction towards the capping domain (Fig. 4d) with residues 199–205 and 207–213 making two short α-helices, and 215–236 forming an extended helical arm (Fig. 4a, d). Only residues Glu181 and Lys182 from P3 make contact with Lys422 and Leu424 from the fingers subdomain of L and a single contact is seen between L and P4 (Fig. 4d). Finally, the helix-forming residues 215–236 of P2, together with two long and mobile helices—presumably belonging to residues 201–217 and 205–231 of P3 and P4, respectively—form a three-helix bundle that is largely exposed to solvent. Consistent with the present observations, previous studies have suggested that residues downstream of the coiled-coil region tend to form helices and constitute a molecular recognition element[11].

Phosphorylation by cellular kinases regulates the transcription modulation activity of VSV-P[13]. Analysis of recombinant HMPV-P using tandem mass spectrometry (Extended Data Fig. 8) showed that P residues targeted by phosphorylation are largely solvent-exposed and changes in phosphorylation should only affect contacts of P with the RNP and M2-1 protein, not complex formation with L (Fig. 4a).

In the *Mononegavirales*, the genome forms a helical nucleocapsid, with RNA bound in a crevice between the two N protein subdomains[14]. P acts as a chaperone by preventing nonspecific aggregation of the N protein in the presence of nonviral RNA, preserving a pool of monomeric RNA-free N protein (N^0) for RNP assembly[15]. The P–N^0 interaction involves an α-helix from HMPV-P (residues 12–28) that inserts into a hydrophobic groove in the CTD of the N^0 protein[15]. Here the entire regions that are N-terminal to the coiled-coil of the four P subunits, including the N-terminal α-helix (P_{NTD}), are disordered. The tetrameric coiled-coil appears to be anchored to the L protein at a location such that these helices, at the tips of flexible polar linkers that span residues 29–135, can bind and deliver N^0 to the RNA exit tunnel for encapsidating nascent RNA (Fig. 4e). The amino acid sequence 29–135 of P (Extended Data Fig. 9) bears physicochemical properties expected of a linker that evolved to ensure the spatial search of nascent RNA exiting L, by N^0 bound to P_{NTD}. In the present structure, the C-terminal regions of P are in nonequivalent positions (Fig. 1f, g). These segments have previously been proposed to dislodge RNA from the RNP complex and feed it into the template tunnel[15] (Fig. 4e). Their structural asymmetry probably reflects different roles for the CTDs of each subunit: whereas the C-terminal end of subunit P2 approaches the RNA template entry tunnel (Fig. 3c) and is well placed to bind N-RNA, the corresponding end of P1 is too distant to have a similar function (Fig. 3a). Moreover, the N terminus of L (residues 8–24) wraps around the template entry tunnel (Fig. 3c). Thus, in addition to the C-terminal helix of P2, the NTD of L could help to displace N to allow naked RNA to feed into the template tunnel (Fig. 3c). In summary, the tetrameric phosphoprotein binds to the monomeric L and fulfils its various adaptor functions by adopting an asymmetric structure (Fig. 4e) through (i) conformational switching, in which each of the four P subunits folds into a different structure, depending on its position in the molecular assembly, and by (ii) spatially confining structurally important N- and C-terminal helical elements (P_{NTD} and P_{CTD}) to the free and RNA-bound N protein, respectively, through long flexible regions that are tethered to the rigidly anchored coiled-coil (Fig. 4e).

While this manuscript was under revision, the structure of RSV L:P was reported[16]. RSV and HMPV L and P proteins have only 48% and 36% overall amino acid identity, respectively, but the structures are similar (r.m.s.d. of 1.34 Å over 1,501 residues). Moreover, the putative RSV-L priming loop retracts to a similar position as seen here for HMPV-L[16]. In RSV, hydrophobic residues that are critical for L:P interactions (L216, L223 and L227 of RSV-P) have been shown to be critical for polymerase function[17]. The corresponding region, residues 251–265 of subunit P1 of the HMPV phosphoprotein, contains several residues that contact L (Fig. 4c). The overall conservation of interactions between L and P within the family *Pneumoviridae* suggests that broad-spectrum inhibitors that disrupt the L:P interface could be effective[16]. We note that

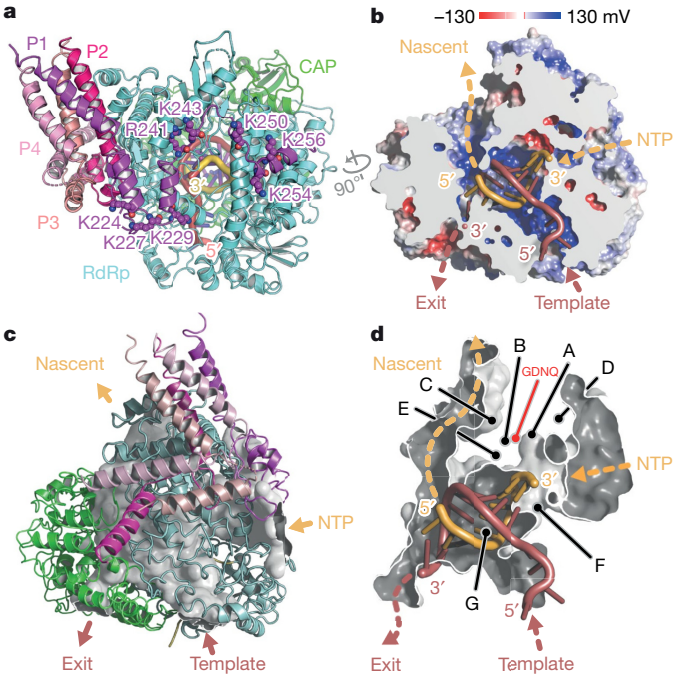


Fig. 3 | Model for RNA elongation by L:P from *Pneumoviridae*. **a**, Model of the HMPV-L:P elongation complex. Template strand, firebrick; nascent strand, orange. Residues from L (Lys307, Arg313 and Arg788) and from the C terminus of subunit P1 (Lys at positions 224, 227, 229, 243, 250, 254 and 256, and Arg241) form a positively charged arch that attracts rNTPs to the rNTP entry tunnel. **b**, Cut-out view of L showing electrostatic surfaces (blue, positive; red, negative) and paths followed by the template and nascent RNA strands. **c**, Tunnels and interior cavities of L depicted as a white surface enclosed by the L:P complex (ribbons). Cut open are various tunnel openings that lead to the exterior. **d**, Magnified view of the tunnel that traverses the L protein. Functional motifs A–G and catalytic residues ^{744}GDNQ^{747} are indicated.

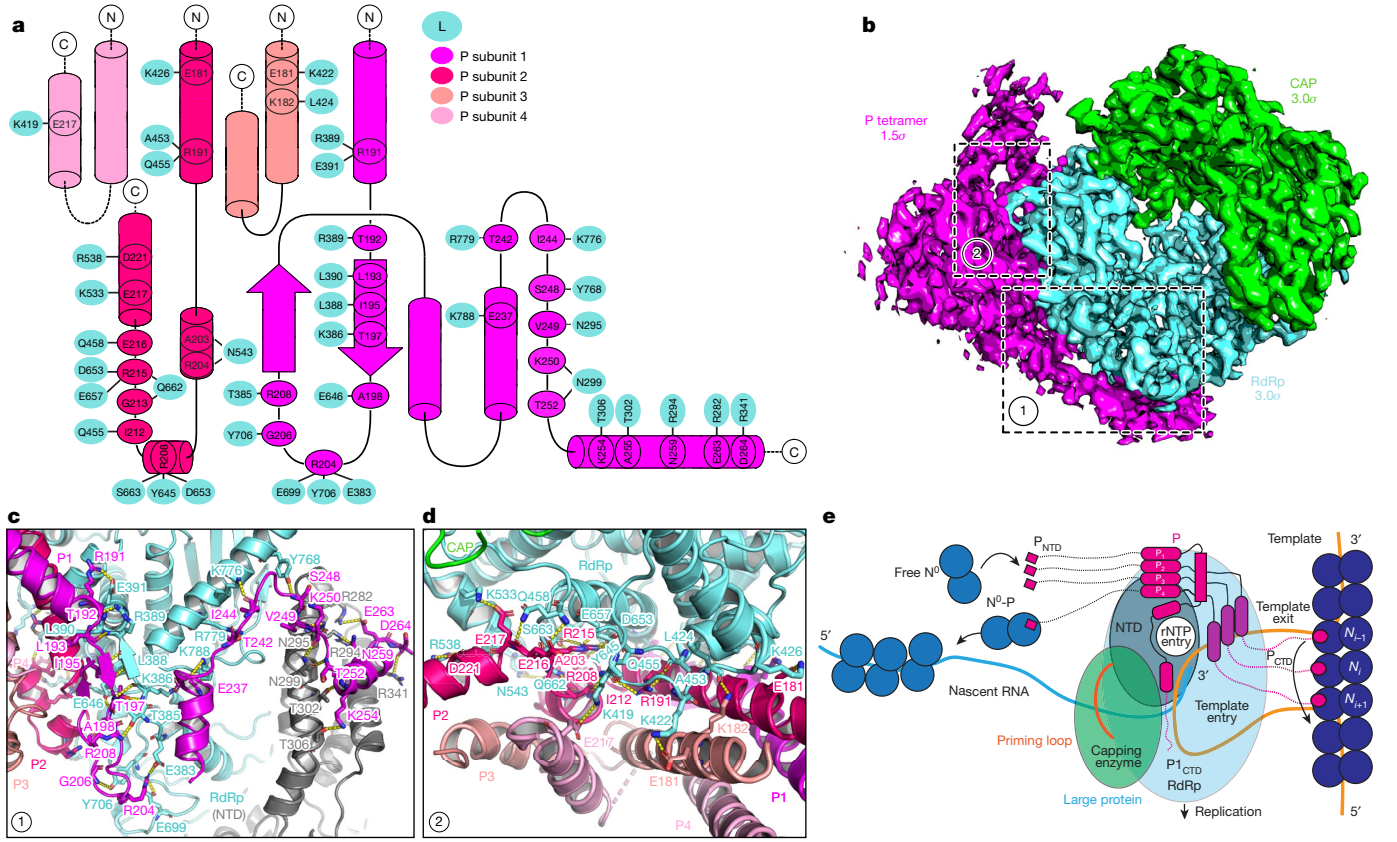


Fig. 4 | The P homotetramer and L:P interactions. a, Interactions between residues from P1 (magenta), P2 (hot pink), P3 (salmon), P4 (pink) of P and the RdRp (cyan). **b**, The electron-density map (RdRp, cyan; capping domain, green; and P, magenta). **c**, A β-hairpin of P1 forms an antiparallel β-sheet with residues 383–391 of the fingers subdomains of L (cyan). Polar contacts between P1 and L are indicated by dashes. **d**, Interactions of the coiled-coil of P and other parts of P2, P3 and P4 with L. **e**, Model for RNA replication by L:P. Genomic viral RNA, extruded from three adjacent N units from the RNP, inserts as a loop of approximately 27 nt into the RNA entry tunnel and back through the template exit tunnel. Genomic RNA displacement is achieved by three P_{CTD} of phosphoprotein subunits P2, P3 and P4 (whereas P1 has a structural function). Concomitantly, four P_{NTD} helices, flexibly linked to the tetramerization domain, bring four new free N^0 subunits towards the nascent RNA exit tunnel for encapsidation.

dimeric P proteins from rabies virus that interact with the C-terminal region of L[18] clearly differ from the tetrameric P proteins of HMPV, which bind to the N-terminal part of L. Whether the mode of interactions between L and P observed here for HMPV and for RSV[16] is also found in more-distant *Mononegavirales* viral families such as the *Filoviridae* requires further structural investigation.

Online content

Any methods, additional references, Nature Research reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41586-019-1759-1.

1. van den Hoogen, B. G. et al. A newly discovered human pneumovirus isolated from young children with respiratory tract disease. *Nat. Med.* **7**, 719–724 (2001).
2. Neubauer, J., Ogino, M., Green, T. J. & Ogino, T. Signature motifs of GDP polyribonucleotidyltransferase, a non-segmented negative strand RNA viral mRNA capping enzyme, domain in the L protein are required for covalent enzyme–pRNA intermediate formation. *Nucleic Acids Res.* **44**, 330–341 (2016).
3. Liang, B. et al. Structure of the L protein of vesicular stomatitis virus from electron cryomicroscopy. *Cell* **162**, 314–327 (2015).
4. Pflug, A., Guilligay, D., Reich, S. & Cusack, S. Structure of influenza A polymerase bound to the viral RNA promoter. *Nature* **516**, 355–360 (2014).
5. Paesen, G. C. et al. X-ray structure and activities of an essential *Mononegavirales* L-protein domain. *Nat. Commun.* **6**, 8749 (2015).
6. Qiu, S., Ogino, M., Luo, M., Ogino, T. & Green, T. J. Structure and function of the N-terminal domain of the vesicular stomatitis virus RNA polymerase. *J. Virol.* **90**, 715–724 (2015).
7. Deval, J. et al. Molecular basis for the selective inhibition of respiratory syncytial virus RNA polymerase by 2′-fluoro-4′chloromethyl-cytidine triphosphate. *PLoS Pathog.* **11**, e1004995 (2015).
8. Tao, Y., Farsetta, D. L., Nibert, M. L. & Harrison, S. C. RNA synthesis in a cage—structural studies of reovirus polymerase λ3. *Cell* **111**, 733–745 (2002).
9. te Velthuis, A. J. W., Robb, N. C., Kapanidis, A. N. & Fodor, E. The role of the priming loop in influenza A virus RNA synthesis. *Nat. Microbiol.* **1**, 16029 (2016).
10. Ogino, M., Gupta, N., Green, T. J. & Ogino, T. A dual-functional priming-capping loop of rhabdoviral RNA polymerases directs terminal de novo initiation and capping intermediate formation. *Nucleic Acids Res.* **47**, 299–309 (2019).
11. Leyrat, C., Renner, M., Harlos, K. & Grimes, J. M. Solution and crystallographic structures of the central region of the phosphoprotein from human metapneumovirus. *PLoS ONE* **8**, e80371 (2013).
12. Blondot, M. L. et al. Structure and functional analysis of the RNA- and viral phosphoprotein-binding domain of respiratory syncytial virus M2-1 protein. *PLoS Pathog.* **8**, e1002734 (2012).
13. Takacs, A. M., Barik, S., Das, T. & Banerjee, A. K. Phosphorylation of specific serine residues within the acidic domain of the phosphoprotein of vesicular stomatitis virus regulates transcription in vitro. *J. Virol.* **66**, 5842–5848 (1992).
14. Tawar, R. G. et al. Crystal structure of a nucleocapsid-like nucleoprotein–RNA complex of respiratory syncytial virus. *Science* **326**, 1279–1283 (2009).
15. Renner, M. et al. Nucleocapsid assembly in pneumoviruses is regulated by conformational switching of the N protein. *eLife* **5**, e12627 (2016).
16. Gilman, M. S. A. et al. Structure of the respiratory syncytial virus polymerase complex. *Cell* **179**, 193–204 (2019).
17. Sourimant, J. et al. Fine mapping and characterization of the L-polymerase-binding domain of the respiratory syncytial virus phosphoprotein. *J. Virol.* **89**, 4421–4433 (2015).
18. Chenik, M., Schnell, M., Conzelmann, K. K. & Blondel, D. Mapping the interacting domains between the rabies virus polymerase and phosphoprotein. *J. Virol.* **72**, 1925–1930 (1998).

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

© The Author(s), under exclusive licence to Springer Nature Limited 2019

Methods

Data reporting

No statistical methods were used to predetermine sample size. The experiments were not randomized and the investigators were not blinded to allocation during experiments and outcome assessment.

Cloning and expression

HMPV L and P genes (from hMPV isolate CAN97-83, GenBank AY297749.1)[19], codon-optimized for expression in *Spodoptera frugiperda* 9 (Sf9) cells, were chemically synthesized and cloned into modified pFastBacDual (Thermo Fischer Scientific) by Bio Basic. HMPV-L (protein sequence, GenBank AAQ67700.1) was inserted downstream of a Strep-tag-II followed by a TEV cleavage sequence. Strep–TEV–HMPV-L was inserted downstream of polyhedrin promoter in pFastBacDual. HMPV-P (protein sequence, GenBank AAQ67693.1) was inserted downstream of an 8×His-tag sequence, followed by a TEV cleavage sequence. Then, 8×His–TEV–HMPV-P was inserted downstream of p10 promoter in pFastBacDual. We mutated pFastBacDual-HMPV-L:P by plasmid PCR to generate pFastBacDual-HMPV-L(D745A):P using forward primer 5′-GACGTGTCCAAGCCTGTGAAGCTGTC-3′ and reverse primer 5′-GATG GACTGGTTCGCGCCGTTGAGC-3′. The mutation was confirmed by sequencing. Generation and isolation of the bacmid of HMPV-L:P and HMPV-L(D745A):P followed the instructions manual of the bac-to-bac expression system (Invitrogen). Viral stocks generated from purified bacmids were amplified and used for protein expression.

Protein purification

The pellet of Sf9 cells was resuspended in 1/10 culture volume in cold HMPV-L:P lysis buffer, 20 mM Na-HEPES, pH 7.4, 300 mM NaCl, 15 mM imidazole, 1 mM $MgCl_2$, 0.5 mM TCEP, and lysed by sonication. Cell debris was pelleted by centrifugation at 50,000*g* for 45 min at 4 °C. Supernatant was incubated with 3 ml pre-equilibrated Ni-NTA beads (Thermo Fisher) per litre of cell pellet for 2 h at 4 °C. The beads were collected by centrifugation at 500*g* for 3 min at 4 °C. The pelleted beads were with 15 mM, 25 mM and 50 mM imidazole with 40, 30 and 20 bed volumes, respectively. Washed beads were transferred into a gravity flow column for elution. HMPV-L:P complex was eluted in HMPV-L:P His-trap elution buffer containing 20 mM Na-HEPES, pH 7.4, 300 mM NaCl, 500 mM imidazole, 1 mM $MgCl_2$, 0.5 mM TCEP, through gravity flow at 4 °C. Eluted fractions were analysed using SDS–PAGE and stained with InstantBlue (Expedeon) to visualize the HMPV L and P proteins (Extended Data Fig. 1a). Fractions containing the HMPV-L:P complex were pooled with the addition of 1/10 (mg ml^{-1}) of TEV and subjected to dialysis against HMPV-L:P lysis buffer with no imidazole at 4 °C overnight. Cleaved HMPV-L:P proteins were incubated with 1 ml Ni-NTA beads at 4 °C for 1 h. Flow-through and wash fractions containing L:P were pooled and subjected to heparin chromatography: a 5-ml HiTrap Heparin (GE Healthcare) column was equilibrated with 20 mM Na-HEPES, pH 7.4, 300 mM NaCl, 1 mM $MgCl_2$, 0.5 mM TCEP, and the protein sample was loaded, washed with the same buffer and eluted using a buffer containing 20 mM Na-HEPES, pH 7.4, 1 mM $MgCl_2$, 1 M NaCl, 0.5 mM TCEP. Fractions that contained pure L:P were pooled, concentrated and loaded onto a Superose 6 increase 10/300 column (GE Healthcare) pre-equilibrated with 20 mM Na-HEPES, pH 7.4, 300 mM NaCl, 1 mM $MgCl_2$, 0.5 mM TCEP. Fractions collected were subjected to SDS–PAGE and instant blue staining. Fractions that contained the L:P complex were pooled and concentrated using Vivaspin15R with a 100-kDa cut-off concentrator (Sartorius) to a concentration of 0.85 mg ml^{-1} for cryo-EM. The final yield for the L:P complex used for cryo-EM was 0.045 mg l^{-1}. During the heparin chromatographic step, a large excess of P was observed to flow through the heparin column. This fraction was further purified by size-exclusion chromatography using a superdex 200 10/300 column (GE Healthcare) pre-equilibrated with 20 mM Na-HEPES, pH 7.4, 300 mM NaCl, 1 mM $MgCl_2$, 0.5 mM TCEP and the pure P was finally concentrated in a Vivaspin15R concentrator (10 kDa cut-off). The yield of free P was 0.23 mg l^{-1}. To obtain protein for the polymerization assay, both the wild-type L:P and L(D745A):P mutant were purified following the same procedure, but without the heparin chromatography step. The L protein in the L:P preparations was quantified by separation on SDS–PAGE and detection using colloidal blue staining. The molar concentration of L:P was determined with respect to L.

RdRp activity assays

PAGE-purified RNA oligoribonucleotides representing residues 1–14 (Fig. 1d) or 1–25 (Extended Data Fig. 2b) of the le sequence were purchased from Sigma and used as a template for in vitro RNA synthesis. Purified recombinant HMPV-L:P at a concentration of 13 nM was incubated with 1 μM RNA oligonucleotide in a total reaction volume of 20 μl in vitro transcription buffer (50 mM Tris-HCl, pH 7.5, 40 mM KCl, 5.0 mM $MgCl_2$, 0.2 U μl^{-1} murine RNase inhibitor, 0.01% (v/v) Triton X-100 and 1 mM DTT). The reactions were pre-equilibrated on ice for 30 min. Transcription reactions were initiated by the addition of rNTPs (1 mM each of rATP, rGTP and rCTP, and 20 μM rUTP) and 20 nM [α-^{32}P] rUTP (3,000 Ci $mmol^{-1}$, Perkin Elmer) or as indicated in the legends and incubated at 30 °C for 3 h. Reactions were terminated by addition of an equal volume of deionized formamide. After addition of proteinase K, the reaction mixtures were further incubated at 37 °C for 20 min. The reaction products were heat-denatured for 10 min at 95 °C and a quarter of the reaction volume was resolved by denaturing urea PAGE (20% acrylamide, 0.5× TBE). Gels were subsequently dried onto a positively charged Hybond N+ nylon membrane (GE Healthcare) and imaged using a phosphor imager. Data are representative of three independent experiments.

The length of the RNA products of the RdRp activity assay was determined by comparison with 3-nt and 25-nt long 5′-triphospho-oligo-RNAs synthesized in vitro using T7 RNA polymerase and [α-^{32}P]rATP incorporation. In brief, annealed complementary DNA oligonucleotides, representing a class II T7 phage promoter followed by the anticipated 25-nt *trailer* RNA product sequence, were incubated with T7 RNA polymerase (New England Biolabs) and [α-^{32}P]rATP according to the manufacturer's instructions. To generate the 3-nt long 5′-triphospho-oligo-RNA marker (5′-pppACG), the reaction was performed under the same experimental conditions except that rGTP was substituted with a 3′-dGTP chain terminator.

To characterize the 3′ extension activity of HMPV polymerase on the RNA template, the le25 template was labelled with *Escherichia coli* poly(A) polymerase (New England Biolabs) in the presence of [α-^{32}P]rATP and excess of the chain-terminating analogue cordycepin 3′-triphosphate (3′-dATP). A quantity of 1 μM RNA was incubated (37 °C, 60 min) with 0.25 U μl^{-1} *E. coli* poly(A) polymerase in reaction buffer (50 mM Tris-HCl (pH 7.9), 250 mM NaCl, 10 mM $MgCl_2$) supplemented with 20 μM 3′-dATP and 100 nM [α-^{32}P]rATP (3,000 Ci $mmol^{-1}$, Perkin Elmer).

Primer elongation reactions contained 2 μM template (containing a 3′ phosphate group), 20 μM primer (containing a 5′-hydroxyl group) and 20 μM each of rATP, rCTP, rGTP and rUTP, with 60 nM [α-^{32}P]rGTP tracer in a buffer containing 50 mM Tris-HCl, pH 7.4, 8 mM $MgCl_2$, 5 mM DTT and 10% glycerol. Primer elongation was initiated by the addition of HMPV-L:P to a concentration of 20 nM in a total reaction volume of 25 μl. Reactions were allowed to proceed for 2 h at 30 °C. After heat inactivation for 3 min at 90 °C, reactions were heat denatured for 5 min at 95 °C in an equal volume of 10 mM EDTA in deionized formamide with 0.02% each bromophenol blue and xylene cyanol dyes. Next, 7 μl of the 50-μl stopped reaction mix was subjected to denaturing urea PAGE (20% acrylamide, 7 M urea, 1× TBE). The gel was dried onto 3MM chromatography paper and the radiolabelled products were detected by phosphor-imaging analysis.

Cryo-EM

We glow-discharged Quantifoil R 1.2/1.3 holey carbon 200-mesh copper grids (Quantifoil Micro Tools) at 20 mA for 45 s. To vitrify the sample, we applied 3.5 µl of the HMPV-L:P complex (both diluted to 0.8 mg ml^{-1}) to each grid, and plunge-froze the sample in liquid ethane using a Vitrobot Mark IV (FEI) after a 4.0-s blotting (offset −2) with filter papers pre-saturated for 15 min under 100% humidity. We screened the sample homogeneity and cryo preparation in an Oxford-style CT3500 cryo-specimen holder (Gatan) on a liquid-nitrogen-cooling Tecnai F20 electron microscope (FEI), operated at 200 kV and equipped with a K2 summit direct detector (Gatan). We screened for thin and clean ice with a maximum distribution of nonoverlapping protein particles. Sample homogeneity was judged by visual observation and by two-dimensional classification (Extended Data Fig. 1). We optimized and prepared two sets of cryo-grids, one with the purified L:P sample and the other with purified L:P supplemented with the free P sample at a 1:0.4 molar ratio.

We automated data acquisition using SerialEM[20] on a K2 summit detector (Gatan) equipped with a Tecnai Polara electron microscope (FEI) operated at 300 kV and at a nominal magnification of 31,000×. The calibrated physical pixel size was 1.24 Å, corresponding to a calibrated magnification of 40,323× on the camera sensor plane. We confirmed beam parallelization by collecting a few videos on carbon after comma-free and objective astigmatism alignments and before the actual data collection; the Thon rings of carbon images consistently extended beyond 2.9–2.7 Å and were of excellent stigmation (Extended Data Fig. 1). For all datasets, we recorded 40 frames per video in super-resolution mode (0.62 Å per pixel) using a dose rate of 9.2 electrons per physical pixel per second (1.24 Å per pixel), and a total exposure of 48 electrons per Å^2. We recorded a total of 7,438 and 1,100 dose-fractionated images for the L:P sample and L:P supplemented with P sample, respectively.

Image processing

We aligned the super-resolution video frames using MotionCor2[21] with dose filtering and 5-by-5 patched sub-frame alignment; the frames were binned over 2-by-2 super-resolution 'pixels' in Fourier space, yielding the physical pixel size of 1.24 Å of the saved micrographs. We binned the motion-corrected micrographs by a factor of 4 (4.96 Å per pixel) and low-pass-filtered to 15 Å for visual examination. We discarded micrographs showing thick ice, abnormal defocus or large carbon areas. For particle-picking and two-dimensional classification, we down-sampled the micrographs to 2.48 Å per pixel in Fourier space using resample_mp.exe[22] to facilitate computation. We manually picked particles for 10 representative micrographs using e2boxer.py[23] and used these boxes to train the auto-picking model in crYOLO[24]. We then used the trained model to automatically pick particles from all micrographs with the crYOLO Phosaurus network. We visually reviewed the results and manually improved the automatically picked boxes for the 10 micrographs, while incorporating missed representative micrographs for the next round training. We repeated the crYOLO training and autopicking procedures until all particles were accurately picked and well-centred for all micrographs. The final L:P dataset contained a total of 5,381,943 particles, and the L:P supplemented with P dataset contained 606,410 particles.

We determined the defocus values and defocus angles of the unbinned (1.24 Å per pixel) micrographs using ctffind4[25]. The Thon ring fitting resolution (reported by ctffind4) of most selected micrographs consistently fell into the 2.9–3.3 Å range (Extended Data Fig. 1), indicating good data quality. We performed two-dimensional classification in Relion[26] until all high-quality two-dimensional class images became apparent (within 50 iterations). We only selected the best two-dimensional classes judged by visual examination and by the resolution and other statistics reported by Relion, retaining 880,724 particles for three-dimensional analysis.

To speed up computation, we started the initial three-dimensional analysis with the data down-sampled to 2.48 Å per pixel. We generated the ab initio model (3 classes) using Relion, and selected the best class as a reference model for further data analysis. We performed one round of mode-3 alignment for the two-dimensional class images against the three-dimensional initial model in frealign (version 9.11)[22], with CTF correction turned off. We then calculated projection matching using frealign (version 9.11)[22] to cross-validate the three-dimensional model and to confirm the two-dimensional class selection as a self-consistency test. We performed three-dimensional classification with 6 classes in Relion until convergence (within 50 iterations). Two classes were clearly of high quality and superimposed well, with a cross-correlation coefficient higher than 99%. We combined these two classes, yielding a total of 437,611 particles. The handedness of the maps was confirmed to be correct as judged by well-defined α-helices. From this point on, we scaled the alignment parameters to 1.24 Å per pixel and used the full resolution data for further analysis. Relion auto-refinement of these particles produced a map with a resolution of 4.0 Å. We further improved the resolution to 3.9 Å after refining the per-particle motion in Relion and to 3.8 Å after beam tilt refinement by treating all particles as one group. Refining the defocus of individual particles did not yield obvious improvement. We confirmed convergence of motion correction by another round of per-particle motion refinement, which yielded no apparent improvement. Nonetheless, the quality of this map at 3.8 Å resolution was excellent, as judged by main-chain continuity and side-chain densities (Extended Data Fig. 1g, h). We further classified the 437,611 particles into 6 classes and discarded 4 junk classes, yielding a total of 302,346 particles. We auto-refined the two good classes individually or combined and again they overlapped nearly perfectly. We then gave the data more weights by adjusting the T number in Relion and subsequently performed masked alignment, while correcting the Fourier shell correlation (FSC) for effects of the solvent mask. We visually examined the results and found out that a T number of 4 produced the best map without apparent signs of over-fitting. This, combined with another round of per-particle motion refinement, beam tilt refinement and per particle defocus refinement, eventually yielded a map at 3.7 Å. In an effort to improve the density associated with P, we generated masks for P and L using Chimera segger[27] and performed multi-body refinement using Relion, with the orientation of P relative to L fixed. We also subtracted the L signal from the two-dimensional images, excised P from the two-dimensional images, and performed three-dimensional classification without alignment. Neither approach improved the P density, perhaps owing to insufficient signal considering the small size of P. The final map was at 3.7 Å resolution and was sharpened with a *B*-factor value of −188.25 Å^2. The procedure is summarized in Extended Data Fig. 3.

We followed the same data-processing paradigm with minor adjustments for the dataset of the L:P preparation supplemented with P. The classification results were consistent with the L:P dataset, leaving 273,574 and 65,398 particles for the good classes after two-dimensional classification and a first round of three-dimensional classification, respectively. The final map after a second round of three-dimensional classification had a resolution of 4.1 Å comprising 29,376 particles and was nearly identical to the map from the L:P dataset. Hence, we mainly used the map from the L:P dataset for map interpretation, except for two less-well-defined C-terminal helices of two P subunits (P3 and P4).

Model building and refinement

We built the atomic model using the graphics programs O[28] and Coot[29]. First, we fitted the RdRp and capping enzymatic domains of the VSV-L model (RCSB Protein Data Bank code (PDB) 5A22) into the map as well as the coiled-coil domain of P (PDB 4BXT) and then used these as a guide to find main-chain connectivity. Local resolution analysis indicated that the majority of the core portion of the complex falls within a resolution range of 3.0–3.6 Å, with excellent main-chain connectivity and

side-chain densities throughout the molecules (Extended Data Fig. 1g, h). The majority of the electron density showed excellent continuity giving confidence for tracing the α-carbon chain. Moreover, in most regions, the clear electron density of the side chains enabled us to unambiguously assign the sequence register. We observed only weak and disordered extra densities accounting for a portion of the CD and CRVI+ domains in maps that were sharpened with lower *B*-factors and/or low-pass-filtered with lower spatial frequency, suggesting that these domains are highly flexible in the complexes that we examined. The appendage of HMPV-L might become fixed with respect to the ring-shaped core domain upon binding to a peptide encompassing residues 35–106 of the VSV-P used in the VSV-L study. Indeed, a recent study of the rabies virus L protein in complex with the counterpart of this P peptide demonstrates such contacts between P and L (J. Horwitz and S. C. Harrison, personal communication). Although there were some stretches with weaker densities for the P protein, long side chains of arginine and lysine residues ensured the correct polarity and register assignment of the sequence. We further confirmed that the sequence was in register by comparing the model with the coiled-coil interface of the known structure of the phosphoprotein tetramerization region (PDB 4BXT). Overall, we traced the entire NTD subdomain and RdRp domain spanning residues 8–892, except for mobile residues 1–7 (predicted to be disordered using the PrDOS prediction server http://prdos.hgc.jp/cgi-bin/top.cgi), 607–625 (603–625 are predicted to be disordered), the complete capping enzyme domain (residues 893–1380) and the four P subunits (residues 169–266, 171–236, 168–219 and 169–193/204–231 of subunits P1, P2, P3 and P4, respectively). Residues 194–203 of subunit P4 are disordered. Whereas the sequences could be assigned unambiguously for P1 and P2 residues as well as for the coiled-coil region of P3 and P4, the sequence assignment for the C-terminal ends of P3 and P4 that are predicted to form helices is less certain. These segments might become fully ordered only in the presence of the ribonucleoprotein (main text). Model building was interspersed with refinement using the real-space refinement option in the package PHENIX[30] (rigid_body, minimization_global, local grid searching and ADP) for the coordinates and *B*-factors. We applied Ramachandran, rotamer and secondary structure restraints throughout the refinement, in addition to the standard geometry and *B*-factor restraints. More than 40 rounds of manual rebuilding and phenix real-space refinement were carried out until we had high confidence of the sequence register, density fitting, model geometry and local chemical environments. We analysed the final model with MolProbity[31] and statistics are given in Extended Data Table 1. We deposited the refined coordinates to wwPDB (6U5O) and the electron density map to EMDB (EMD-20651).

Modelling of the polymerase–RNA elongation complex

We fitted a rotavirus in situ elongation complex (PDB 6OJ3) to the HMPV-L:P complex by superimposing the CA atoms of the five functional polymerase motifs in the program O[28] and refined the alignment using the O command lsq_improve. The relatively long RNA enabled us to easily identify the NTP entry, template entry, template exit and transcript exit tunnels in the HMPV-L structure. Modelling using elongation complexes of other polymerases such as influenza virus PB1 (PDB 6QCT)[32], the HCV NS5 (PDB 4WTG)[33] and reovirus λ3 elongation complex (PDB 1N35) confirmed the results.

Tunnel identification

We identified the tunnels in the structure using voidoo, mama and mapman of the Uppsala Software Factory[34]. We tested values for the probe radius in the range of 1.4–2.0 Å, and 1.8 Å yielded the most accurate tunnels. We generated pseudo-atoms to define the range of the calculation and displayed the tunnel along with the models in PyMOL (https://pymol.org/2/). We performed object-specific clipping of the tunnel mask using PovRay (http://www.povray.org) to display its interior along with the modelled RNA.

Tandem mass spectrometry

We collected and analysed tandem mass spectrometry data (Extended Data Fig. 8) for both the purified L:P complex and free P following a standard protocol[35]. In brief, we digested 2 μg of each sample overnight at 37 °C with trypsin (1:100 enzyme:sample molar ratio). Digests were carried out in 200 mM EPPS pH 8.5 in the presence of 2% acetonitrile (v/v) with trypsin (Promega, V5111). After peptide purification by reverse-phase C18 chromatography, 1 μg of peptides was separated with a 4-h acetonitrile gradient by high-performance liquid chromatography before injection. Data were collected using an Orbitrap Lumos mass spectrometer (Thermo Fisher Scientific) coupled to a Proxeon EASY-nLC 1200 liquid chromatography system (Thermo Fisher Scientific). The 75-μm inner-diameter capillary column used was packed with C_{18} resin (Accucore, 2.6 μm). The scan sequence started with collection of an MS^1 spectrum (Orbitrap analysis; resolution 120,000; mass range 400–1400 Th). MS^2 analysis followed collision-induced dissociation (CE = 35) with a maximum ion-injection time of 250 ms and an isolation window of 1.2 Da. Neutral loss of phosphate was taken into account by using a multi-stage activation method with a neutral loss of 97.9763 Da. Peptides were searched with a SEQUEST-based algorithm against an insect database containing L and P protein sequences and also against common contaminants. Searches were performed with a target decoy database strategy and a false-discovery rate of 1% set for peptide-spectrum matches followed by linear discriminant analysis filtering and a final collapsed protein-level false-discovery rate of 1%. We dynamically searched for oxidized methionine (+15.9949146221 Da) and phospho-serine, -threonine and -tyrosine (+79.9663304104 Da). Phosphorylation site localization used a modified analysis of variance (ANOVA) score (ModScore). We repeated the experiments twice and the results were consistent.

RSV priming loop experiments

Mutations in codon-optimized RSV L (strain A2) were generated in a fragment of L open-reading frame subcloned into a pGEM T easy vector (Promega). The W1262A substitution was generated by site-directed mutagenesis by Genscript, the P1261A, P1274A, Y1276A substitutions were generated using Q5 site-directed mutagenesis (New England Biolabs). RSV(P1261A) was generated using forward primer 5′-ACCTACTAAGGCCTGGGTCGG-3′ and reverse primer 5′-CCCCTCTCGCCCCTGGTC-3′. RSV(P1274A) was generated using forward primer 5′-CCAGGAAAAAAAGACCATGGCCGTCTACAACAGG-3′ and reverse primer 5′-GTGGAGGAGCCGACCCAGGGCTTAGTAGG-3′. RSV(Y1276A) was generated using forward primer 5′-CATGCCCGTCGCCAACAGGCAG-3′ and reverse primer 5′-GTCTTTTTTTCCTGGGTG-3′. The mutated open-reading frame fragments were sequenced and then substituted in the codon-optimized L open-reading frame contained in a T7 expression vector (pTM1) by restriction digestion and ligation. All plasmids were sequenced to confirm the presence of mutations, and to ensure that no other changes were introduced during PCR-amplification steps. BSR-T7 cells, a BHK-derived cell line that constitutively expresses T7 polymerase[36] (provided by K. Conzelmann), in six-well dishes were transfected with pTM1 plasmids expressing the following proteins: RSV (A2 strain) N (400 ng per well), P (200 ng per well), M2-1 (100 ng per well) and L (or L mutant; 100 ng per well), and a plasmid expressing a minigenome RNA containing nt 1–36 of the *trailer* promoter at its 3′ end (200 ng per well). The minigenome was limited to synthesizing only the first strand of replication by the lack of a promoter-specific sequence at the 3′ end of the replicative intermediate. The cells were transfected using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions. After transfectionf for 40–48 h, cells were collected and RNA was purified using TRIzol, according to the manufacturer's instructions, except that the RNA was subjected to a further round of purification by extraction with phenol-chloroform, followed by ethanol precipitation. RNA products

initiated within the *trailer* promoter were analysed by primer extension. RNA representing one-tenth of a well of cells was subjected to reverse transcription at 37 °C using the Sensiscript RT kit (Qiagen) and a radiolabelled primer (5′-TACGAGATATTAGTTTTTGAGAC-3′). Half of the primer extension reaction was subjected to electrophoresis in 8% polyacrylamide gels containing 7 M urea in 1× TBE. Radiolabelled oligonucleotides corresponding in sequence to cDNA representing initiation from positions +1 and +3 of the *trailer* promoter were used as markers (+1, 5′-TACGAGATATTAGTTTTTGAGACTTTTTTTCTCGT-3′; +3, 5′-TACGAGATATTAGTTTTTGAGACTTTTTTTCTC-3′). The same RNA samples with RNA representing one-tenth of a well of cells were also subjected to denaturing gel electrophoresis on a 1.5% formaldehyde–agarose gel and transferred to nitrocellulose by northern blotting. The input minigenome template was detected using a radiolabelled strand-specific riboprobe. Primer extension products were visualized by autoradiography and the primer extension products and input minigenome were quantified using phosphor-imaging analysis. The levels of input minigenome were used to normalize for transfection efficiency. The data were normalized to the mean of the two wild-type samples included in each experiment. With the caveat that the sample sizes had an n of 3, statistical analysis using ANOVA showed that the levels of +1 and +3 RNA products for the P1261A and W1262A L variants were significantly different from those of wild-type L ($P = 0.002$ and $P = 0.003$, respectively, for +1; $P = 0.004$ for both variants for +3), whereas the levels of RNA produced from P1274A and Y1276A L variants were not statistically different ($P = 0.0385$ and $P = 0.549$, respectively, for +1; $P = 0.227$ and $P = 0.518$, respectively, for +3).

Figure preparation

We prepared the figures and videos using PyMOL (Schrödinger), Chimera[27], Python matplotlib, gnuplot, CCP4 TopDraw[37] and PovRay (http://www.povray.org). Multiple sequence alignment was performed with Clustal Omega[38] (https://www.ebi.ac.uk/Tools/msa/clustalo/) and the alignment displayed with ESPript[39] using previous studies[40–43] to assign the sequence motifs.

Reporting summary

Further information on research design is available in the Nature Research Reporting Summary linked to this paper.

Data availability

Structure coordinates are available from PDB 6U5O and the electron density map from the Electron Microscopy Data Bank (EMDB) under accession code EMD-20651. All other data generated or analysed in this study are available from the corresponding authors upon reasonable request.

19. Biacchesi, S. et al. Genetic diversity between human metapneumovirus subgroups. *Virology* **315**, 1–9 (2003).
20. Mastronarde, D. N. Automated electron microscope tomography using robust prediction of specimen movements. *J. Struct. Biol.* **152**, 36–51 (2005).
21. Zheng, S. Q. et al. MotionCor2: anisotropic correction of beam-induced motion for improved cryo-electron microscopy. *Nat. Methods* **14**, 331–332 (2017).
22. Lyumkis, D., Brilot, A. F., Theobald, D. L. & Grigorieff, N. Likelihood-based classification of cryo-EM images using FREALIGN. *J. Struct. Biol.* **183**, 377–388 (2013).
23. Tang, G. et al. EMAN2: an extensible image processing suite for electron microscopy. *J. Struct. Biol.* **157**, 38–46 (2007).
24. Wagner, T. et al. SPHIRE-crYOLO is a fast and accurate fully automated particle picker for cryo-EM. Commun. *Biol.* **2**, 218 (2019).
25. Rohou, A. & Grigorieff, N. CTFFIND4: fast and accurate defocus estimation from electron micrographs. *J. Struct. Biol.* **192**, 216–221 (2015).
26. Scheres, S. H. W. RELION: implementation of a Bayesian approach to cryo-EM structure determination. *J. Struct. Biol.* **180**, 519–530 (2012).
27. Pettersen, E. F. et al. UCSF Chimera—a visualization system for exploratory research and analysis. *J. Comput. Chem.* **25**, 1605–1612 (2004).
28. Jones, T. A., Zou, J.-Y., Cowan, S. W. & Kjeldgaard, M. Improved methods for building protein models in electron density maps and the location of errors in these models. *Acta Crystallogr. A* **47**, 110–119 (1991).
29. Emsley, P., Lohkamp, B., Scott, W. G. & Cowtan, K. Features and development of Coot. *Acta Crystallogr. D* **66**, 486–501 (2010).
30. Adams, P. D. et al. PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallogr. D* **66**, 213–221 (2010).
31. Chen, V. B. et al. MolProbity: all-atom structure validation for macromolecular crystallography. *Acta Crystallogr. D* **66**, 12–21 (2010).
32. Kouba, T., Drncová, P. & Cusack, S. Structural snapshots of actively transcribing influenza polymerase. *Nat. Struct. Mol. Biol.* **26**, 460–470 (2019).
33. Appleby, T. C. et al. Structural basis for RNA replication by the hepatitis C virus polymerase. *Science* **347**, 771–775 (2015).
34. Kleywegt, G. J. & Jones, T. A. Detection, delineation, measurement and display of cavities in macromolecular structures. *Acta Crystallogr. D* **50**, 178–185 (1994).
35. Kalocsay, M. APEX peroxidase-catalyzed proximity labeling and multiplexed quantitative proteomics. *Methods Mol. Biol.* **2008**, 41–55 (2019).
36. Buchholz, U. J., Finke, S. & Conzelmann, K. K. Generation of bovine respiratory syncytial virus (BRSV) from cDNA: BRSV NS2 is not essential for virus replication in tissue culture, and the human RSV leader region acts as a functional BRSV genome promoter. *J. Virol.* **73**, 251–259 (1999).
37. Bond, C. S. TopDraw: a sketchpad for protein structure topology cartoons. *Bioinformatics* **19**, 311–312 (2003).
38. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
39. Robert, X. & Gouet, P. Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.* **42**, W320–W324 (2014).
40. Poch, O., Blumberg, B. M., Bougueleret, L. & Tordo, N. Sequence comparison of five polymerases (L proteins) of unsegmented negative-strand RNA viruses: theoretical assignment of functional domains. *J. Gen. Virol.* **71**, 1153–1162 (1990).
41. van den Hoogen, B. G., Bestebroer, T. M., Osterhaus, A. D. M. E. & Fouchier, R. A. M. Analysis of the genomic sequence of a human metapneumovirus. *Virology* **295**, 119–132 (2002).
42. Bruenn, J. A. A structural and primary sequence comparison of the viral RNA-dependent RNA polymerases. *Nucleic Acids Res.* **31**, 1821–1829 (2003).
43. Lang, D. M., Zemla, A. T. & Zhou, C. L. E. Highly similar structural frames link the template tunnel and NTP entry tunnel to the exterior surface in RNA-dependent RNA polymerases. *Nucleic Acids Res.* **41**, 1464–1482 (2013).
44. Kucukelbir, A., Sigworth, F. J. & Tagare, H. D. Quantifying the local resolution of cryo-EM density maps. *Nat. Methods* **11**, 63–65 (2014).
45. Mason, S. W. et al. Interaction between human respiratory syncytial virus (RSV) M2-1 and P proteins is required for reconstitution of M2-1-dependent RSV minigenome activity. *J. Virol.* **77**, 10670–10676 (2003).
46. Asenjo, A., Calvo, E. & Villanueva, N. Phosphorylation of human respiratory syncytial virus P protein at threonine 108 controls its interaction with the M2-1 protein in the viral RNA polymerase complex. *J. Gen. Virol.* **87**, 3637–3642 (2006).

Acknowledgements We thank S. C. Harrison for support, encouragement, suggestions and reading of the manuscript and acknowledge funding from NIH grant AI089618 (to S. C. Harrison). We thank Z. Li for cryo-EM support, M. Luo and M. Liao for suggestions and help with cryo-EM data acquisition, Y. Liu for suggestions on image-processing strategies, S. Jenni, S. M. Hinshaw and S. P. Lim for suggestions and critiques, and A. Shareef for insights regarding HMPV RNA synthesis assays. The J.L. laboratory was supported by Academic Research Grant Tier 3 (MOE2016-T3-1-003) from the Singapore Minister of Education (MOE) and grant NRF2016NRF-CRP001-063. X.Q. was supported by the Singapore-MIT Alliance for Research and Technology. The R.F. laboratory was supported by NIH grant AI113321.

Author contributions J.P., X.Q., R.F. and J.L. conceived the study. X.Q. and T.H.Y. expressed and purified HMPV-L:P. J.P. collected and processed cryo-EM data. J.P. and J.L. built and refined the structure. S.L. carried out the polymerase activity assays. M.K. performed phosphorylation studies. J.P. and A.E.S. analysed data and prepared figures. H.J. and X.Q. prepared the L mutant. T.C. and S.N. performed priming loop assays and minigenome studies. B.L. performed the primer elongation assay. J.P., X.Q., R.F. and J.L. wrote the paper with input from all authors.

Competing interests The authors declare no competing interests.

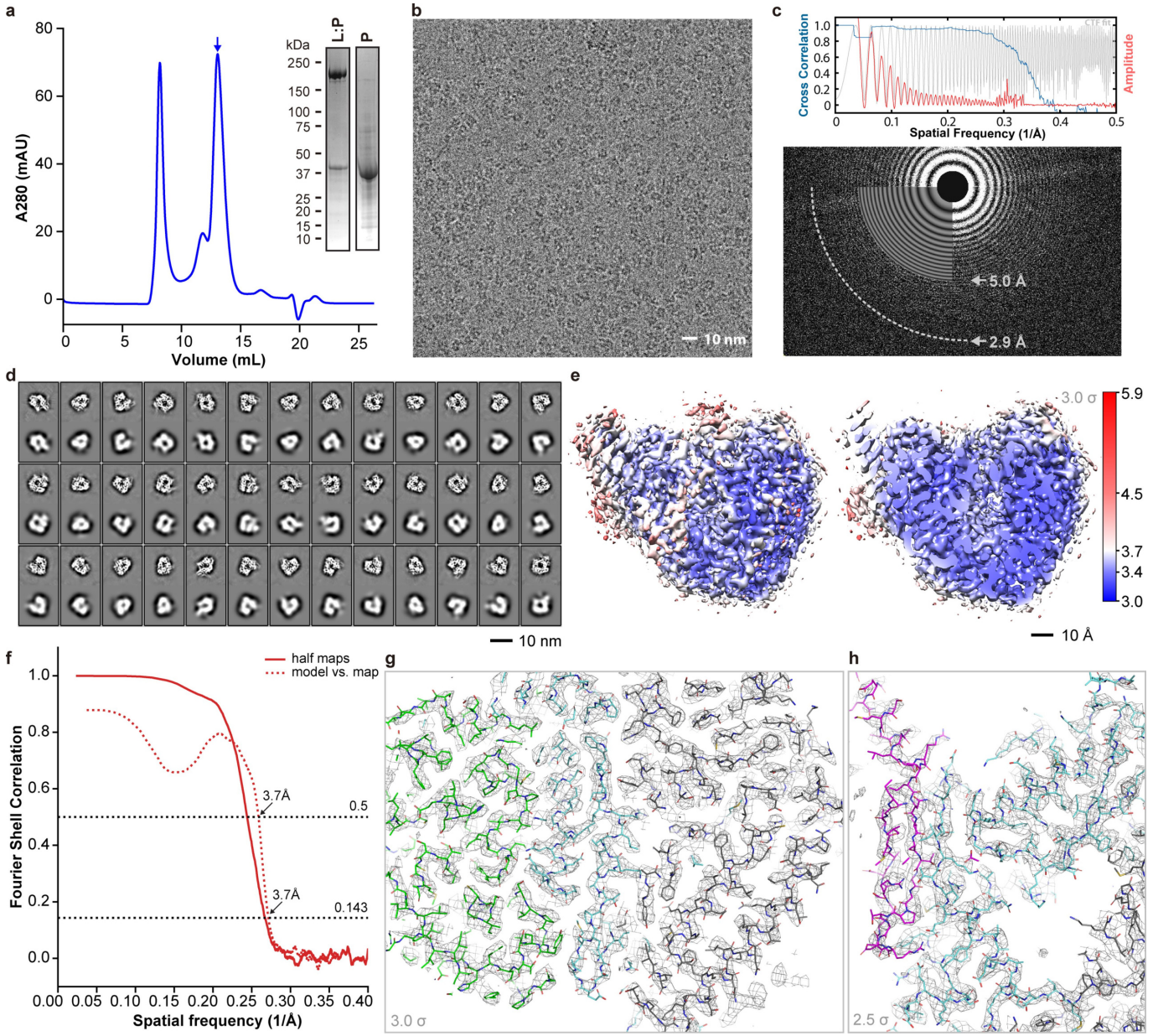
Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41586-019-1759-1.

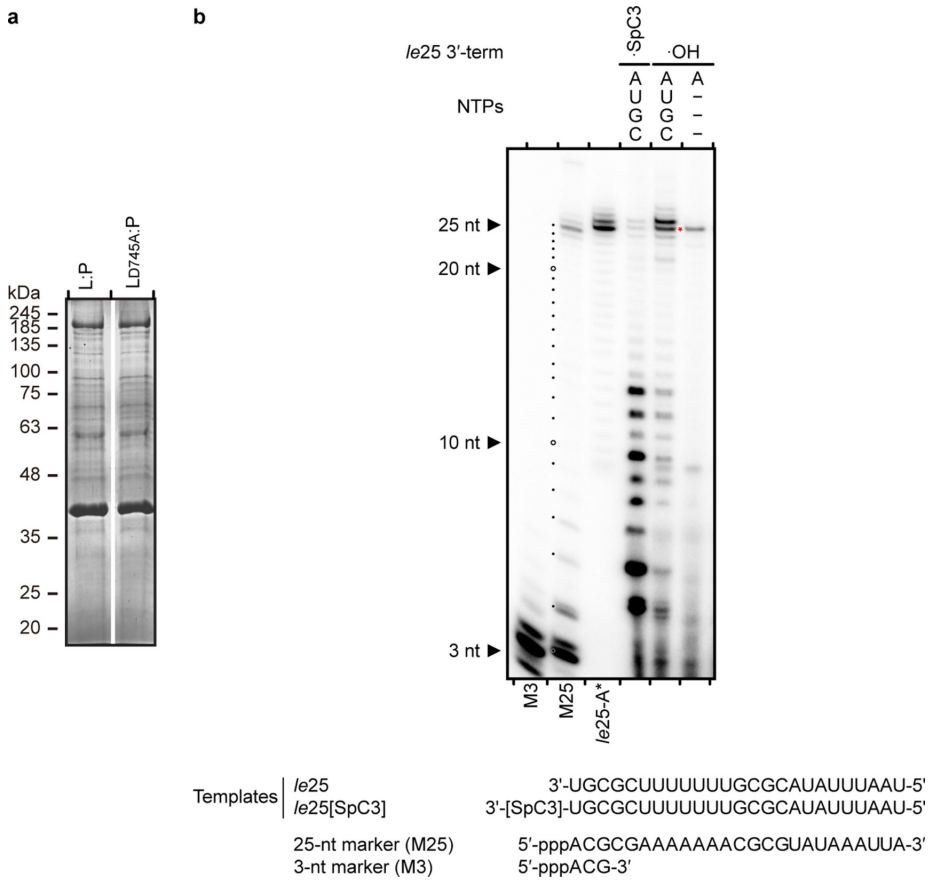
Correspondence and requests for materials should be addressed to J.P., R.F. or J.L.

Peer review information *Nature* thanks Tomoaki Ogino, Aartjan te Velthuis and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

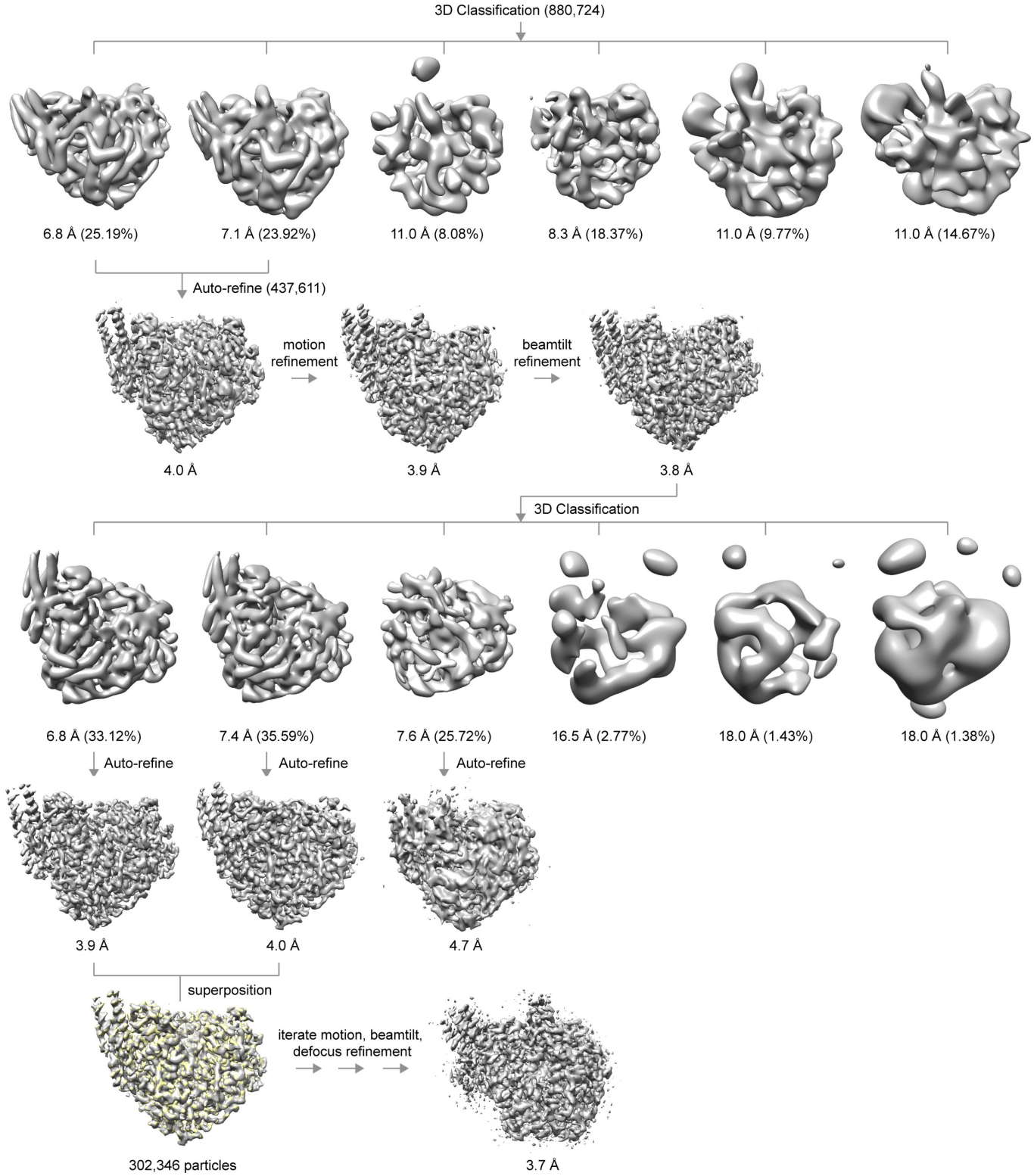
Reprints and permissions information is available at http://www.nature.com/reprints.



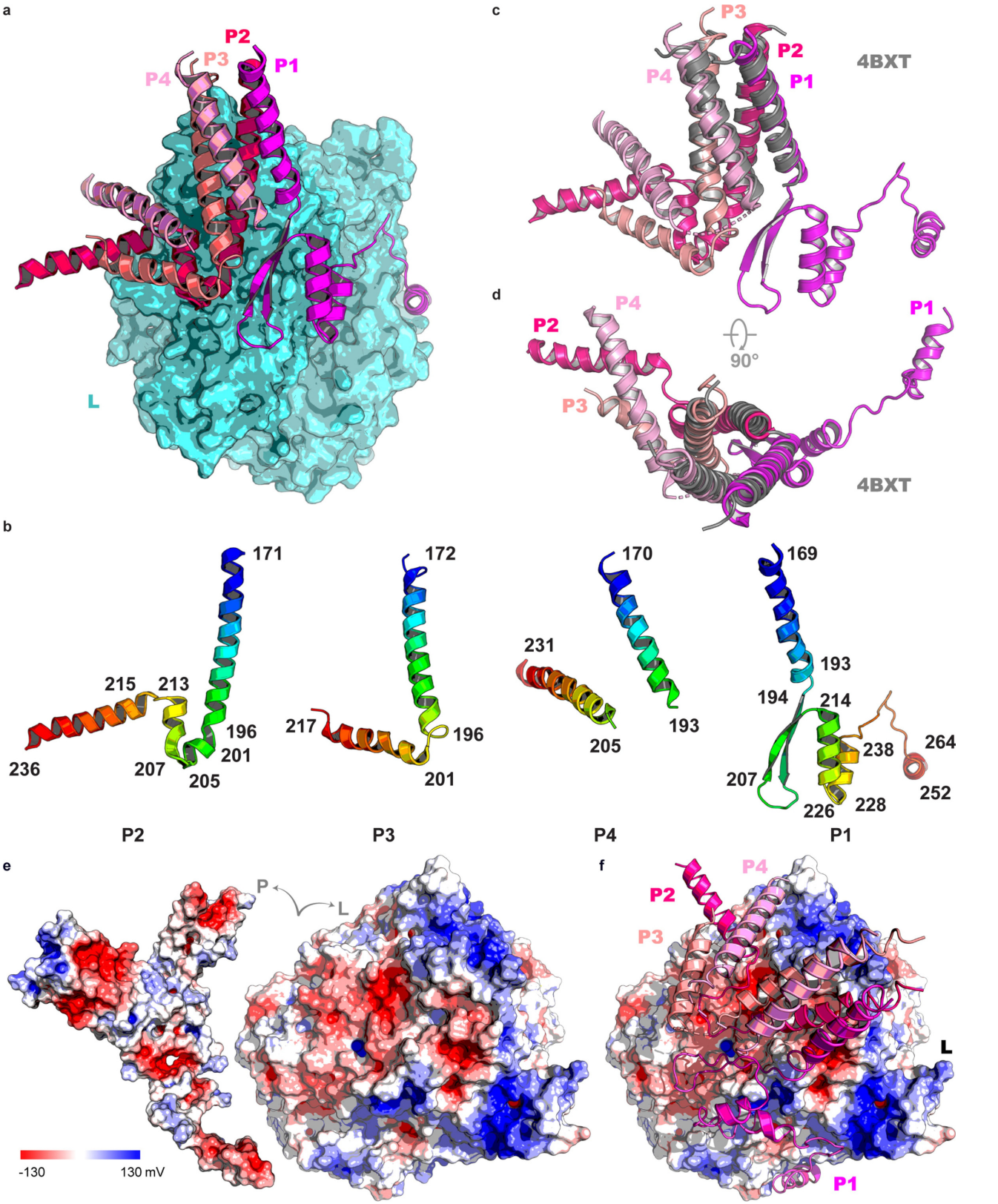
Extended Data Fig. 1 | Purification of HMPV-L:P and structure determination using cryo-EM. a, Representative size-exclusion chromatogram of the L:P complex (these experiments were repeated more than five times). Fractions indicated by an arrow were collected and concentrated to 0.85 mg ml^{-1} and used for cryo-EM analysis. Inset, SDS–PAGE followed by Coomassie blue staining of the purified samples. Free P protein separated from the L:P complex by heparin chromatography is also shown. For gel source data, see Supplementary Fig. 1. **b**, Raw micrograph of HMPV-L:P particles recorded in vitreous ice. Scale bar, 10 nm. **c**, Power spectrum of the image shown in **b**. We limited the high resolution for fitting to a spatial frequency of 1/5.0 Å, and 1/2.9 Å marks the highest spacing to which CTF rings were successfully fit. **d**, Two-dimensional classes and 'self-consistency check' for the cryo-EM three-dimensional reconstruction. For each pair of boxes (top and bottom) across the three rows, the top image shows one two-dimensional class average and the bottom image shows the corresponding projection from the initial three-dimensional model. **e**, Local resolution of the cryo-EM density map. Variations in local resolution are colour-coded from blue (3.0 Å) to red (5.9 Å), computed with Resmap[44]. **f**, FSC of the cryo-EM map as a function of the spatial frequency. The gold-standard resolution is 3.7 Å based on the FSC = 0.143 criterion, consistent with the model-to-map correlation (0.5 criterion). **g**, Example of the electron-density map that enabled model building. The region shown is at an interface between the RdRp and capping domain. The map is shown as a grey mesh, contoured at a level of 3σ. The atomic model is shown as sticks with residues from RdRp coloured in cyan (NTD in grey) and in green for the capping domain. **h**, The region shown is the three-stranded β-sheet at the interface between the RdRp (cyan sticks) and the phosphoprotein (magenta sticks). The map is shown as a grey mesh, contoured at a level of 2.5σ. We observed a nearly identical structure of the L:P complex in a reconstruction obtained by premixing the L:P complex with fully phosphorylated P, indicating that potential exchange of P affected neither the formation nor the structure of the L:P complex.



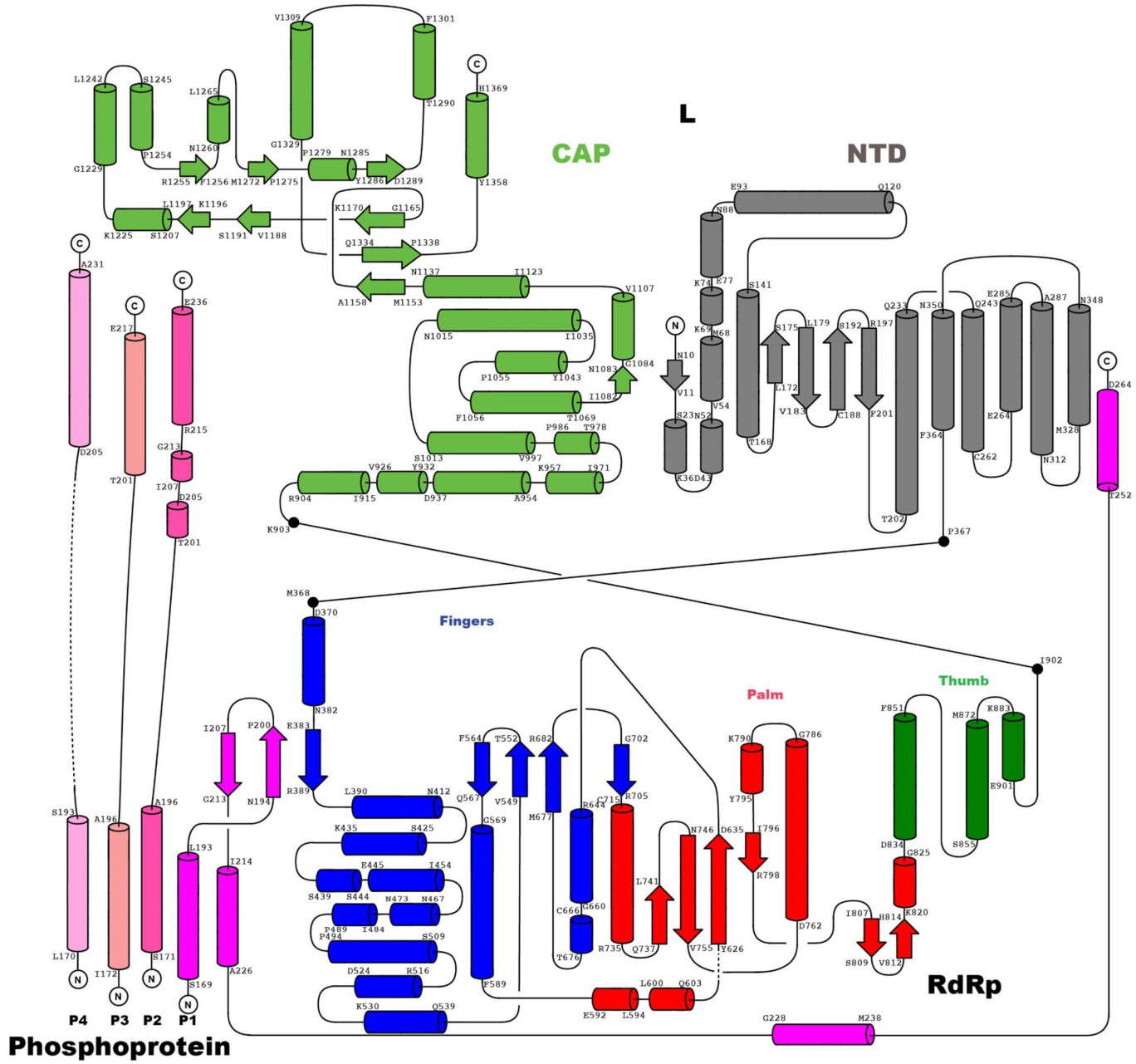
Extended Data Fig. 2 | RdRp activity assay. a, SDS–PAGE of HMPV wild-type L:P, L(D745A):P purified for RdRp activity assays. Proteins were purified by metal affinity, TEV cleavage of the His-tag followed by reverse His-tag affinity purification and size-exclusion chromatography. **b**, Analysis of the 3′ extension activity of HMPV polymerase using the le25 RNA template. Reactions were performed with rNTPs (0.5 mM each of rUTP, rGTP and rCTP), 20 µM rATP and 20 nM [α-^{32}P]rATP. When a 3′-modified le25 (le25[SpC3], three-carbon spacer group linked to the 3′ extremity) was used as a template, synthesis of products longer than 25 nt was greatly reduced compared to le25. When only [α-^{32}P]rATP and no other rNTP was supplied, only a product with a length longer than 25 nt was observed. This result shows that the L:P complex was capable of modifying the 3′ terminus of the template, in addition to engaging in de novo initiation at the promoter. The radiolabelled RNA products were visualized by phosphor imaging. Data are representative of three independent experiments. For gel source data, see Supplementary Fig. 1.



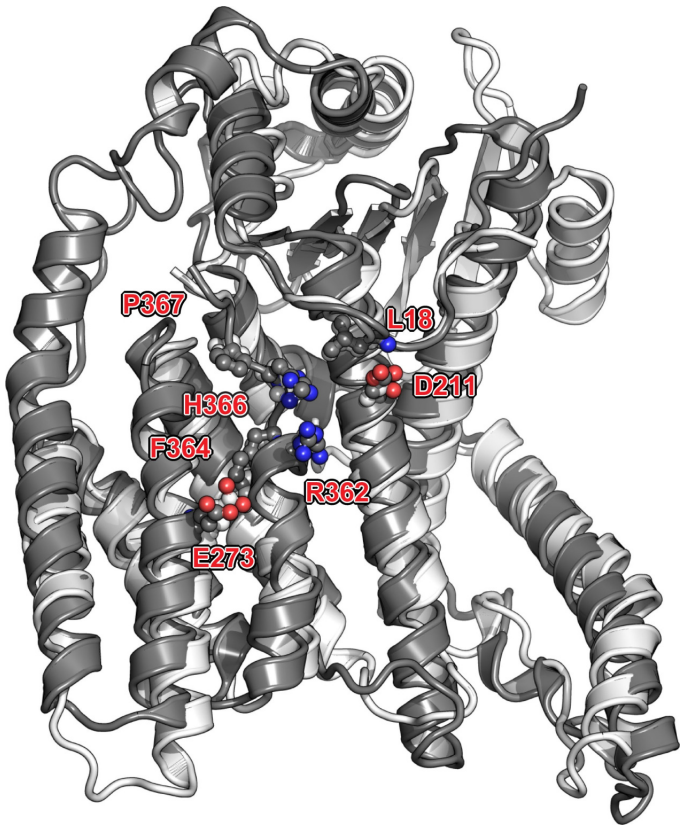
Extended Data Fig. 3 | Flow-chart depicting structure determination using cryo-EM. Further information is provided in the Methods.



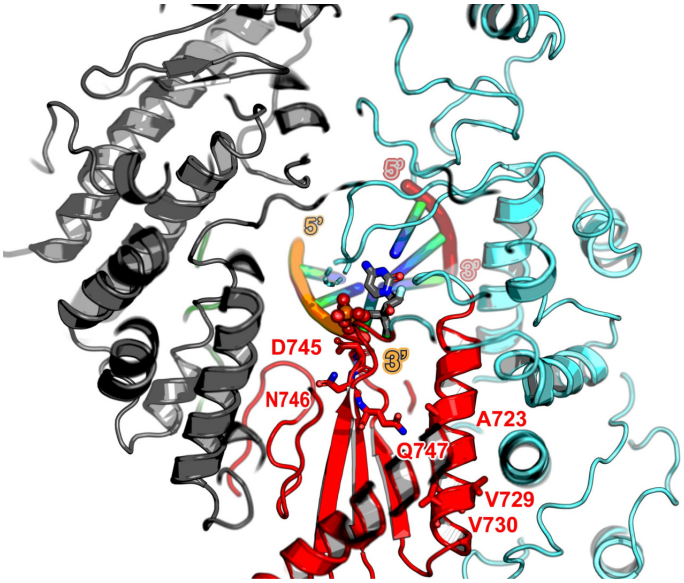
Extended Data Fig. 4 | Phosphoprotein tetramer in complex with L. a, The L protein (cyan) is represented as a molecular surface and the tetrameric P protein subunits are represented as ribbons, following the colour codes in Figs. 1–3. **b**, Structures adopted by the four individual P subunits bound to L, coloured as a blue-to-red rainbow from the N- to the C-terminal end. Secondary structures boundaries are noted for each subunit. **c**, **d**, Superposition of the tetramerization helices in the context of the L:P complex and the free P protein. Structures are represented as coloured ribbons with the free phosphoprotein coiled-coil (PDB 4BXT) coloured in grey and the four P subunits reported in this study coloured according to Fig. 1. The r.m.s.d. of the superimposition is 1.13 Å over 88 α-carbon atoms. **e**, View of the complex in which L and P have been pulled apart to display electrostatic surfaces. **f**, Overall view of the L:P complex with P shown as ribbons and L as electrostatic surface. The P tetramer consists of subunits P1 (magenta), P2 (hot pink), P3 (salmon) and P4 (pink).



Extended Data Fig. 5 | Topology of the L:P complex. Topological depiction of the secondary structure elements of L and P. Helices are depicted as tubes and strands as arrows. The colour code is the same as in Fig. 1. The RdRp domain and its subdomains and the capping domain are coloured as in Figs. 1, 2: NTD in grey, fingers in blue, palm in red, thumb in dark green and CAP in green. The four subunits of the phosphoprotein P1, P2, P3 and P4 are coloured as in Fig. 1. Secondary structure boundaries are indicated.

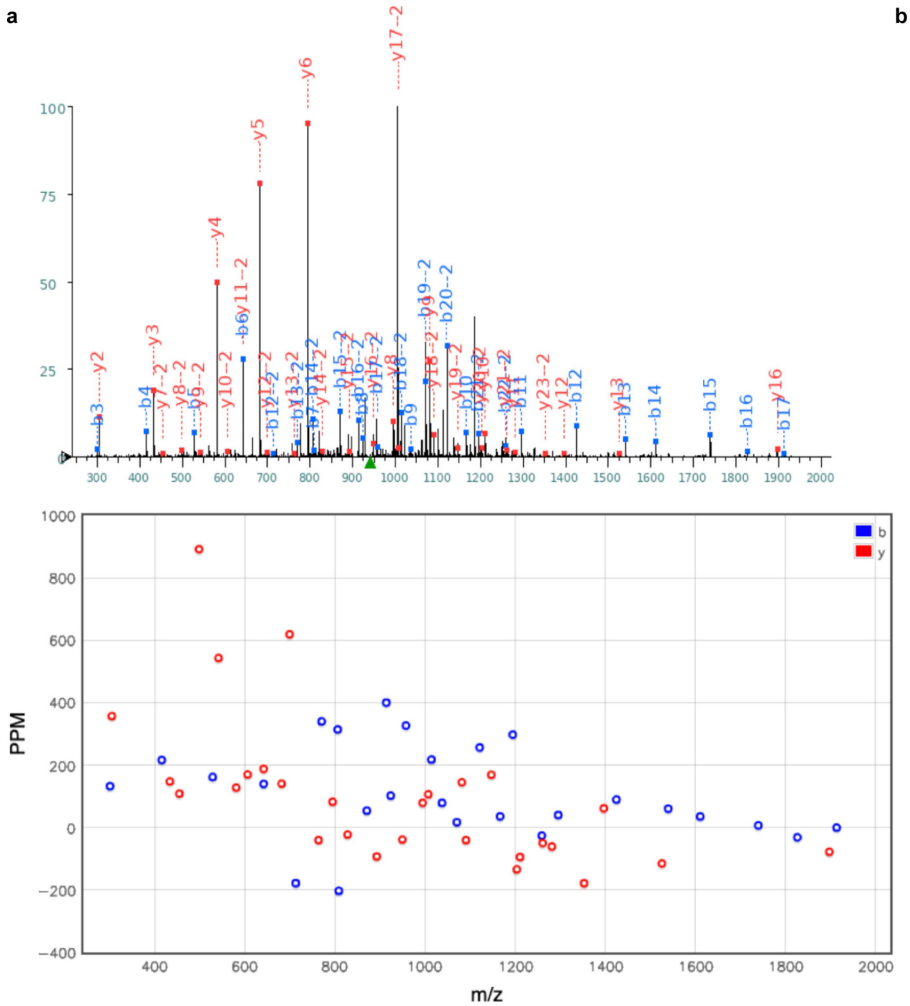


Extended Data Fig. 6 | View of the NTD. NTD is displayed as grey ribbons (colours are as in Fig. 2a). The evolutionary conserved residues clustered near the rNTP entry tunnel play a part in transcription and are represented as sticks and are labelled. The equivalent region of the VSV-L superimposed on HMPV-L is coloured in lighter grey.



Extended Data Fig. 7 | Model for an elongation complex stalled by the addition of ALS-8176 5′-triphosphate. ALS-8176 5′-triphosphate is a nucleoside triphosphate analogue against RSV and HMPV currently in phase 2 clinical trials. The ^{744}GDNQ^{747} catalytic motif and the positions (A789V, L795I and I796V) for which mutations conferred resistance (identified by passaging RSV) are mapped onto the HMPV-L structure (respectively corresponding to A723, V729 and V730) and displayed as sticks. These conservative mutations probably affect inhibitor binding by inducing a slight repositioning of the helix, owing to altered hydrophobic contacts with neighbouring helices. The sequence alignment is shown in Supplementary Fig. 2. the protein is coloured according to Fig. 2a, and the template and nascent RNA strands according to Fig. 3a.

a



b

Sequence

DALDLLS#DNEEEDAESSILTFEER

Predicted Fragmentation Pattern

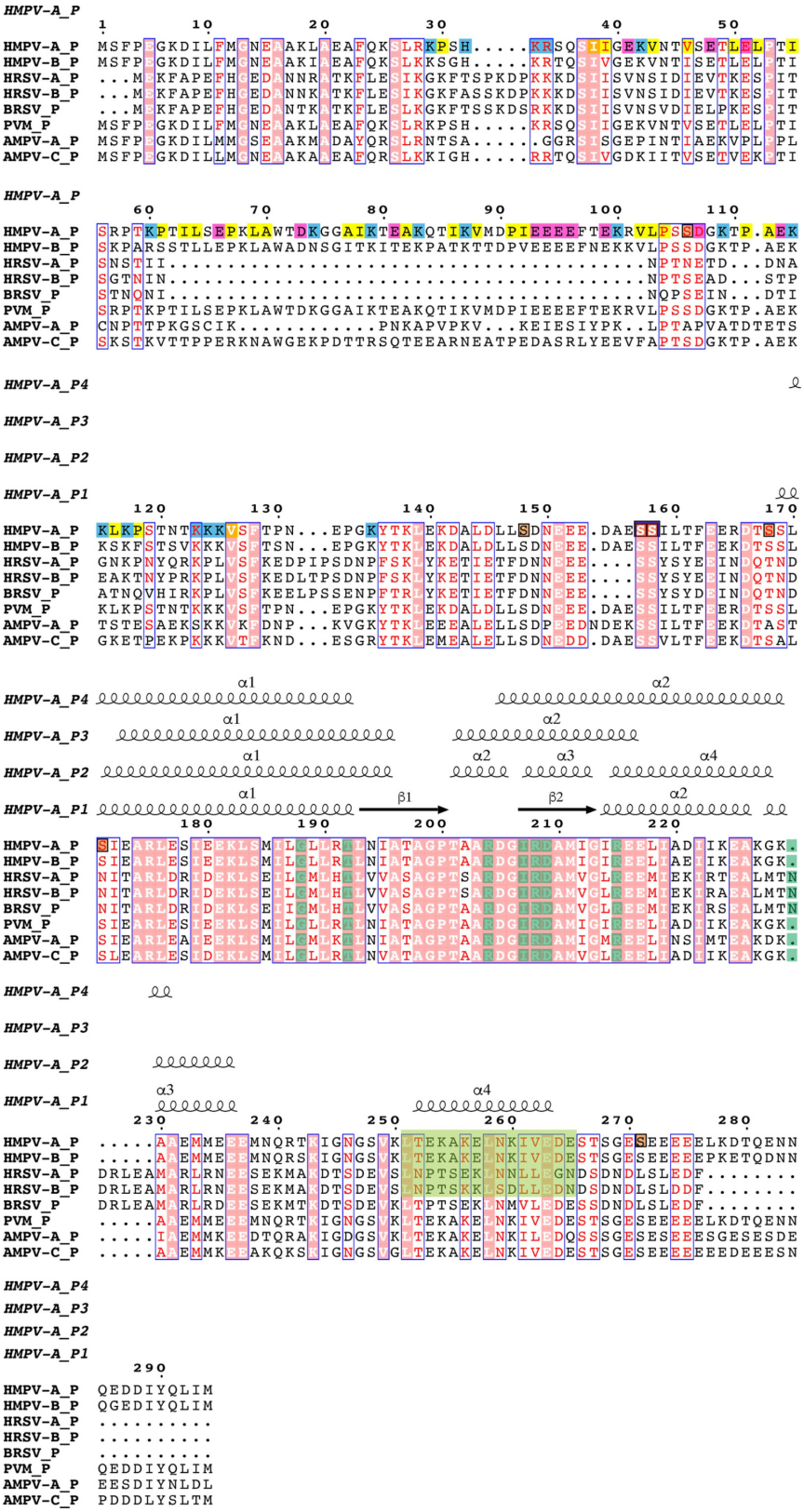
+1

Seq	#	b: Δ Error	b	y	y: Δ Error	+1
D	1	---	116.034	---	---	24
A	2	---	187.071	2705.171	---	23
L	3	132.153	300.155	2634.134	---	22
D	4	216.062	415.182	2521.050	---	21
L	5	162.102	528.266	2406.023	---	20
L	6	139.584	641.350	2292.939	---	19
S#	7	-202.637	808.349	2179.855	---	18
D	8	102.074	923.376	2012.857	---	17
N	9	79.423	1037.419	1897.830	-77.866	16
E	10	35.801	1166.461	1783.787	---	15
E	11	40.251	1295.504	1654.744	---	14
E	12	89.561	1424.546	1525.702	-115.146	13
D	13	60.140	1539.573	1396.659	61.609	12
A	14	35.735	1610.611	1281.632	-60.899	11
E	15	6.848	1739.653	1210.595	-94.120	10
S	16	-31.261	1826.685	1081.552	144.855	9
S	17	-0.074	1913.717	994.520	79.100	8
I	18	---	2026.801	907.488	---	7
L	19	---	2139.885	794.404	82.743	6
T	20	---	2240.933	681.320	140.656	5
F	21	---	2388.001	580.273	128.360	4
E	22	---	2517.044	433.204	147.548	3
E	23	---	2646.087	304.162	356.606	2
R	24	---	---	175.119	---	1

+2

Seq	#	b: Δ Error	b	y	y: Δ Error	+1
D	1	---	58.521	---	---	24
A	2	---	94.039	1353.089	-178.348	23
L	3	---	150.581	1317.571	---	22
D	4	---	208.095	1261.029	-49.681	21
L	5	---	264.637	1203.515	-133.988	20
L	6	---	321.179	1146.973	168.925	19
S#	7	---	404.678	1090.431	-40.435	18
D	8	---	462.192	1006.932	106.318	17
N	9	---	519.213	949.419	-38.387	16
E	10	---	583.734	892.397	-92.987	15
E	11	---	648.256	827.876	-22.599	14
E	12	-177.914	712.777	763.354	-39.867	13
D	13	339.646	770.290	698.833	619.051	12
A	14	314.080	805.809	641.320	187.552	11
E	15	53.785	870.330	605.801	169.846	10
S	16	400.149	913.846	541.280	543.212	9
S	17	326.644	957.362	497.764	891.670	8
I	18	217.724	1013.904	454.248	108.388	7
L	19	16.531	1070.446	397.706	---	6
T	20	256.786	1120.970	341.164	---	5
F	21	297.748	1194.504	290.640	---	4
E	22	-25.889	1259.026	217.106	---	3
E	23	---	1323.547	152.584	---	2
R	24	---	---	88.063	---	1

Extended Data Fig. 8 | MS^2 spectrum of the Ser^{148} P phosphopeptide. a, One MS^2 spectrum used for identification of the phosphorylated P peptide ^{142}DALDLLS#DNEEEDAESSILTFEER is displayed. Tandem mass spectrum (top) and deviation (bottom) enabled detection of phosphorylation (indicated by a hash symbol) at site Ser^{148}. Peptides fragmented from the N terminus (b-fragments) and C terminus (y-fragments) are coloured in blue and red, respectively. **b**, The y and b ion series *m/z* identified in the spectrum shown in **a** and their deviation from theoretical *m/z* are shown in the table. The present pattern of phosphorylation agrees with observations showing that phosphorylation of the peptide comprising residues 100–120 of RSV-P[45]—in particular, phosphorylation of Thr108[46], which corresponds to Ser148 of HMPV-P (Extended Data Fig. 9)—controls its interaction with the M2-1 protein.



Extended Data Fig. 9 | See next page for caption.

Extended Data Fig. 9 | Structure-based sequence alignment of the phosphoprotein from HMPV (labelled HMPV-A, strain CAN97-83) and other known pneumoviral P proteins. HMPV-B, human metapneumovirus subgroup B; HRSV-A and -B, human respiratory syncytial virus subgroup A and B, respectively; BRSV, bovine respiratory syncytial virus; PVM, pneumonia virus of mice; AMPV-A and AMPV-C, avian metapneumovirus subgroup A and C, respectively. The following sequence accession codes from GenBank were used for the alignment: HMPV-A, AAQ67693.1 (used in this study); HMPV-B, AAQ67684.1; HRSV-A, AAX23990.1; HRSV-B, AAR14262.1; BRSV, AAL49395.1; PVM, AAW79177.1; AMPV-A, AAT68644.1; and AMPV-C, AAT86110.1. The secondary structure of HMPV-P subunit P1 (this study) is displayed above the alignment. Phosphorylation sites are highlighted in brown. Positively charged residues of HMPV_P are shaded in blue, negatively charged residues in magenta and hydrophobic residues 29–135 in yellow. The conserved region containing hydrophobic residues critical for L:P interactions is highlighted in green. Structural alignment of P from HMPV and RSV[16] showed similar overall tetramer organization. However, differences are observed in subunit P1 with an r.m.s.d. of 2.24 Å over 82 residues. Although P is generally more mobile with weaker densities and higher *B*-factors compared to L, the region following the β-hairpin (residues 175–215 in HMPV) does adopt a slightly different conformation compared to RSV P1. Subunit P3 has an r.m.s.d. of 1.94 Å over 45 residues, owing to a slightly tilted C-terminal helix compared to RSV. Subunit P2 is most similar with an r.m.s.d. of 0.92 Å over 56 residues. Subunit P4 has an r.m.s.d. of 1.33 Å over 47 residues. The eight residues of HRSV-P, which are crucial for the interaction with HRSV-L and substitutions of which impair viral replication, are shaded in dark green (data from a previous study[16]). With the exception of Asn189 (HRSV-P), for which a deletion is present in HMPV-P, these residues are conserved in HMPV-P and other known pneumoviral P proteins.

Extended Data Table 1 | Cryo-EM data collection, structure refinement and model statistics

	HMPV L:P complex (EMDB-20651) (PDB 6U5O)
Data collection and processing	
Magnification (nominal/calibrated)	31,000/40,322.58
Voltage (kV)	300
Electron exposure (e–/Å^2)	48
Defocus range (μm)	1.5-3.0
Pixel size (Å)	1.24
Symmetry imposed	C1
Initial particle images (no.)	5,381,943
Final particle images (no.)	302,346
Map resolution (Å)	3.7
FSC threshold	0.143
Map resolution range (Å)	∞-3.7
Refinement[a]	
Initial model used (PDB code)	-
Model resolution (Å)	3.7
FSC threshold	0.5
Model resolution range (Å)	∞-3.7
Map sharpening *B* factor (Å^2)	-188.25
Model composition	
Non-hydrogen atoms	12,977
Protein residues	1,623
Ligands	-
B factors (Å^2)	
Protein	15.62 (5.27-52.43)
Ligand	-
R.m.s. deviations	
Bond lengths (Å)	0.004
Bond angles (°)	0.717
Validation[b]	
MolProbity score[c]	1.31 (98th percentile)
Clashscore	2.13
Poor rotamers (%)	0.00
Ramachandran plot	
Favored (%)	100.00
Allowed (%)	0.00
Disallowed (%)	0.00

[a]Refinement statistics were obtained with program Phenix (phenix.real_space_refinement)[30] unless otherwise noted.

[b]Molprobity score was obtained from the Molprobity online server (100th percentile is the best among structures within the resolution specified) and the current structure is in the 98th percentile (*n* = 27,675, 0–99 Å).

[c]The assignment of Pro1041 as a *cis*-proline is supported by electron density showing a hydrogen bond between the carbonyl oxygen atom of Thr1040 and the N atom of Gln1353.

nature research

Corresponding author(s): Pan, Junhua
Lescar, Julien

Last updated by author(s): Oct 11, 2019

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed
☒	☐ The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☒	☐ A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐ The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐ A description of all covariates tested
☒	☐ A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐ A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐ For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐ For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐ For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐ Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	SerialEM (v3.7.1-64bit)
Data analysis	Please see the methods section for details. MotionCor2 (v1.2.1), crYOLO (v1.2.3), ctffind (v4.1.9), EMAN2 (v2.2), Relion (v3.0.4), Frealign (v9.11), ResMap (v1.1.4), O (v14.0.0), Coot (v0.8.8), Phenix (v1.16-3549), CCP4 (v7.0), Buster (20190301), MolProbity (v3), voidoo (v3.3.2), mama (v6.1.9), mapman (v7.8.5), PyMOL (v2.3.2), Chimera (v1.13.1), Pov-Ray (v3.8), matplotlib (v3.1.1), gnuplot (v5.2).

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

cryo-EM maps and coordinates have been deposited and released (EMDB access code: EMD-20651 and PDB code: 6U5O). Scans of raw gels deposited to Supplemental figures associated with this manuscript. All other data generated or analyzed during this study are available from the corresponding authors upon reasonable request.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	Statistical methods were not needed to pre-determine sample size for the biochemical and structural studies in this work. We collected multiple independent cryoEM datasets for structural analysis; all other experiments were performed multiple times with similar results.
Data exclusions	All data were analyzed.
Replication	All biochemical and mass spectrometry experiments were performed multiple times and resulted nearly identical results (Please see methods section for details). Multiple cryoEM datasets were collected and analyzed, yielding very similar results. Evidence of replicaton also included similar results obtained by different data processing strategies, by using different softwares, by results shown in the data processing flowchart (Extended Data Figure 3), and by processing different subsets (Please see the methods section for details).
Randomization	Experimental groups were not needed for this work, and hence randomization was irrelavant.
Blinding	The investigators were blinded to group allocation during data acquisition and analysis because groups were not necessary for this work.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants
☒	☐ Clinical data

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Author Correction: Tumour lineage shapes BRCA-mediated phenotypes

https://doi.org/10.1038/s41586-019-1839-2

Correction to: *Nature* https://doi.org/10.1038/s41586-019-1382-1

Published online 10 July 2019

Philip Jonsson, Chaitanya Bandlamudi, Michael L. Cheng, Preethi Srinivasan, Shweta S. Chavan, Noah D. Friedman, Ezra Y. Rosen, Allison L. Richards, Nancy Bouvier, S. Duygu Selcuklu, Craig M. Bielski, Wassim Abida, Diana Mandelker, Ozge Birsoy, Liying Zhang, Ahmet Zehir, Mark T. A. Donoghue, José Baselga, Kenneth Offit, Howard I. Scher, Eileen M. O'Reilly, Zsofia K. Stadler, Nikolaus Schultz, Nicholas D. Socci, Agnes Viale, Marc Ladanyi, Mark E. Robson, David M. Hyman, Michael F. Berger, David B. Solit & Barry S. Taylor

In this Letter, the *y*-axis label of Fig. 4d should read 'Event-free survival (%)' rather than 'Overall survival (%)', as correctly described in the figure legend, main text and Methods. In addition, in the 'Clinical and treatment data' section of the Methods, the fourth sentence should read: "Treatment data were curated for the patients receiving one or more lines of either PARP inhibitor or immune checkpoint blockade therapy via free-text search **of medical and pharmacy records followed by expert review in a subset of cases, as follows**." instead of: "Treatment data were curated for the patients receiving one or more lines of either PARP inhibitor or immune checkpoint blockade therapy via free-text search and expert review of medical records.". Neither of these changes affects the data, results or conclusions of the Letter, and the original Letter has been corrected online.

Author Correction: Superluminal motion of a relativistic jet in the neutron-star merger GW170817

https://doi.org/10.1038/s41586-019-1860-5

Correction to: *Nature* https://doi.org/10.1038/s41586-018-0486-3

Published online 05 September 2018.

K. P. Mooley, A. T. Deller, O. Gottlieb, E. Nakar, G. Hallinan, S. Bourke, D. A. Frail, A. Horesh, A. Corsi & K. Hotokezaka

In Extended Data Fig. 1 of this Letter, panel a should have been marked b and vice versa. This error has been corrected online.

Work

Your story

Send your careers story to: naturecareerseditor@nature.com



THINGLASS/GETTY

A BIOLOGIST TALKS TO A PHYSICIST

Working alongside people from different fields can help in improving communication skills and identifying gaps in knowledge. **By Kenneth S. Kosik**

As a physician–scientist, many of my colleagues were surprised when I moved my laboratory from the Boston Longwood campus at Brigham and Women's Hospital in Massachusetts to the University of California, Santa Barbara (UCSB), where there is neither a medical school nor a university-affiliated hospital.

More than a few e-mails arrived — some expressed puzzlement, some surprise, some had a wink, but all were punctuated with an exclamation mark.

When I made the move ten years ago, I had wanted to shift my biomedical work closer to the interests of physical-science researchers, to discover broadly applicable principles within the framework of biology, and to grasp the multilayered complexity hidden in nearly every question posed by biologists.

The physical sciences are not lacking in Boston. In fact, they are world class. But my lab was on the 'medical' side of the Charles River, practically an ocean away from the physical-science labs at Harvard University in Cambridge, Massachusetts, on the other side. So, engaging scientists in physics and chemistry, or in computer science and engineering, was challenging — especially for a medical doctor trained rather narrowly in molecular and cellular biology.

My goal was simply to open a conversation and possibly a collaboration with physicists, not to become one. As a relatively small

© 2020 Springer Nature Limited. All rights reserved.

institution with a distinguished faculty in both the physical sciences and engineering, UCSB was, I felt, an ideal place to wade into this territory. Over the ensuing decade, this risky move has resulted in my sharing many graduate students and postdoctoral fellows with computer-science, engineering, physics and chemistry faculty members. These collaborations have greatly broadened the science in all my publications.

I devised a few simple rules to help the biologist in me to cross the divide between the life and physical sciences. In learning to talk to physicists, I discovered that I can communicate better with everyone and clarify for myself what I do and do not understand in my own field.

Understand what 'I do not understand' means. When physicists say they do not understand something that you have said about biology, it's possible that you do not understand that topic either.

'Understanding' operates at different planes in different disciplines, and when a physicist seeks understanding, what they hope to grasp might differ from the knowledge that a biologist seeks.

For example, a biologist understands gene transcription by identifying specific transcription factors, their binding sites, the role of RNA polymerase and the genes that get activated.

For the physicist, these crucial facets of transcription — specific gene names and binding sites — are extraneous details. Instead, among the questions they consider important are the probability distributions associated with attracting transcription complexes to specific sites and the quantification of the forces involved in this process.

Clearly, questions of this nature are of keen interest to biologists once we sidestep our love affair with our favourite gene.

Seek common ground. When a physicist says they do not understand an aspect of biology, they are not requesting a 'biology 101' explanation. In my experience, when physicists ask a biology question, they want to apply the thinking of physics to biology; specifically, they are searching for universal, mathematical explanations.

Physicists move away from settled questions. In biology, much less seems settled. Emphasizing what you know is less interesting than saying what you need to learn.

Many of the missing pieces in biology are quantitative details, such as the absolute copy numbers of a protein or an RNA in a single cell that kinetically mediates some function, and how the cell keeps track of so many regulatory dials. Do cells perform regular maintenance on a parts-replacement calendar, as is done for aeroplanes, or is damage the only trigger for replacing parts? Does damage tend to occur according to a clock? Identifying these unknowns in biology is more stimulating than recounting textbook-level information.



SONIA FERNANDEZ

Neuroscientist Kenneth S. Kosik.

> **"The fastest way to make a physicist's eyes glaze over is to recite biological jargon."**

Recognize the posture of false modesty. I have found that physicists often display a false modesty regarding their knowledge of biology. A physicist will say biology is much more difficult than physics because there is a lot to memorize. They might say: you know an encyclopaedia of detail, and all I know are a few equations.

Keep in mind the maths shortfall in biology. For most biological phenomena, we don't have precise equations — unlike in physics. This is not to say that we don't have maths, but our field needs a lot more detailed quantification. This lack is the Achilles heel of biology, and affects even the concepts we use every day.

It is a rare biologist who can explain the negative binomial used in RNA sequencing, parameterize organelle shape and distance scales when assessing active transport and diffusion, or interpret the phase diagrams that correspond to biomolecular condensates — a surging interest in biology.

Deep computational knowledge is a huge asset for biologists, but many have not had the opportunity to hone these skills or even be introduced to them. In my experience, opening a conversation and sharing students with physical scientists is step one towards bridging this gap.

Don't be flummoxed by physicists' maths. In discussing their own work, physicists will often reach for a formula. After they write the equation and stare at it as if pondering a Mark Rothko painting, they might proffer an explanation. You can have a productive conversation without the bread-and-butter basics of physics, such as Hamiltonians or modelling an Ising phase transition. Bear in mind that physicists can be more sensitive about the maths than are mathematicians, who are not bound by reality and see the world as a constraint.

Albert Einstein said: "As far as the laws of mathematics refer to reality, they are not certain; and as far as they are certain, they do not refer to reality." The biologist must have substantial comfort with uncertainty. Compared with the good predictions from theory in physics, prediction in vast facets of biology, such as evolution, is more like stock-market cycles — discernible only in retrospect.

Scale matters. Consider this matter of perspective: most of the time, what a physicist does is much smaller or much bigger, or much colder or much hotter, than anything we biologists do.

© 2020 Springer Nature Limited. All rights reserved.

Biologists should know the sizes of macromolecules — the organelles and the cells that they work with — and first consider how they scale (or do not scale) with systems derived from physics.

Consider precision. Physicists require much greater certitude to draw a conclusion. They might be a bit smug when you claim significance with a *P* value of less than 0.05. For them, attaining a result with a *P* value of 0.05 is like the chance of hitting a barn door with a baseball: you can't miss it.

Avoid jargon. The fastest way to make a physicist's eyes glaze over is to recite biological jargon, such as gene lists and pathways. Come to think of it, it's the fastest way to make anyone's eyes glaze over.

Skip some details. Physicists do not want to know about all your controls. They want the concepts, and are willing to assume you have performed the appropriate control experiments and have chosen the correct methods. If you're doing a 'chalk talk' (illustrated presentation) for physicists, be sure to emphasize your ideas more than detailed data and techniques. Physicists make judgements on the basis of clear, cogent, compelling ideas.

Manage expectations. Physicists expect you to build your own equipment. If this topic comes up with a theorist, feel free to ignore it. If it comes up with an experimentalist, you will have to concede that you paid a huge mark-up for off-the-shelf equipment.

Understand optimization versus the 'good-enough' principle. Biological processes carry evolutionary baggage, and therefore arrive at solutions that might not be optimal. If, for example, you were designing a machine that walks, it might not be best to model it on human walking, which is not necessarily the most efficient way to get somewhere. In the absence of evolution, the physical world is often assumed to operate according to optimized parameters. By contrast, biology operates in what has been called a Waddington terrain: processes function well even when they are not as efficient as possible — as long as they are still good enough to work.

Consider a physicist's perspective on reductionism. Reductionist approaches to complex biological systems do not necessarily mean reducing those systems to small units. Systems approaches to biology that involve large numbers of variables can be inherently interesting to physicists because they conceptually resemble problems in statistical mechanics that relate microscopic and macroscopic properties with very large degrees of freedom — and exact solutions are not possible. If a biologist can frame a problem in these terms, it will probably engage a physicist.

Physicists laugh a lot. Not only is the humour of physicists arcane, but almost anything unexpected can provide a jocular moment. Theirs are the ultimate inside jokes, which are often not obviously funny. But laugh along anyway — even if you don't find the humour, they won't know the difference.

The future

To conclude, let me say this path is not for everyone. Knowing yourself is the crucial element. If you are uncomfortable engaging with specialists in fields in which you are not trained, or if you don't want to delve into topics that might not immediately bear a clear connection to your established domain, then the course I took is not for you. However, if you are not only comfortable with this approach, but require stimulation from outside your immediate realm of investigation, there is nothing better than to look in unexpected places.

Kenneth S. Kosik is a professor of neuroscience at the University of California, Santa Barbara.

A PHYSICIST TALKS TO A BIOLOGIST

The fastest way to understand a new discipline is to embed yourself in it. **By Sarah Bohndiek**

It's amazing how many of biologist Ken Kosik's points in his accompanying careers column (page 281) resonate with my experience as a physicist working in the opposite direction. I enjoyed my training in physics largely because it didn't involve rote learning (needed by biologists to get that encyclopaedic knowledge) and it allowed me to derive much of the material from a few key equations and principles.

In his column, Ken notes that when a physicist says they do not understand something in biology, they are not requesting a lengthy 'biology 101' explanation. But I do sometimes want a biology 101 refresher so I can gather a deeper understanding of the biological problem while I think about how to address it.

Ken also talks about being uncomfortable with uncertainty. Personally, I can be quite comfortable with uncertainty when I can take steps to control it; it's just that, in biological systems, I normally have to accept that I can't.

A deeper understanding

The PhD I received from University College London in 2008 was in radiation physics, with a focus on evaluating the potential of CMOS image sensors (the silicon behind most smartphone cameras) for application in medical X-ray-diffraction studies.

Most of this research took place in a dark room, where I examined the optical response of the image sensors, but because I was in a department of medical physics and bioengineering, I was exposed to a range of biomedical challenges — including cancer detection.

As I progressed through my studies, I realized that I wanted to gain a deeper understanding of cancer itself and how we might exploit its biology so we could use non-invasive imaging to find the disease.

I hadn't studied biology since I was 16, so I started looking for postdoctoral positions in 'friendly' environments — biophysics, bioengineering and so on. But I soon realized that, to truly understand cancer, I needed to break out of my comfort zone and immerse myself in that environment.

My first postdoc was in Kevin Brindle's biochemistry laboratory at the University of Cambridge, UK. I learnt how to hold a pipette, run a magnetic resonance imaging system, conduct animal experiments and design biology studies.

In return, I fixed temperamental magnets, wrote hardware-control code and supervised some of the lab's more physical-sciences-oriented projects. After three years, I had a much greater understanding of cancer biochemistry and how to conduct *in vivo* imaging studies, but I was missing exposure to clinical applications and connection to my previous research home in optics.

My second postdoc, with Sam Gambhir in the department of radiology at Stanford University, California, took me back to my roots. I explored new projects in optics, trying to make imaging with optics faster and cheaper than before and opening new horizons in fields such as endoscopy.

As part of Stanford's Molecular Imaging Program, I saw teams build new technologies in the

© 2020 Springer Nature Limited. All rights reserved.

lab and take them through to first-in-human studies. I was inspired by Sam's dedication to revolutionizing early cancer detection and his incredible passion for transforming technology into clinical trials. That passion still drives me in my current role at the University of Cambridge, UK, where the mission of my interdisciplinary team is to advance our understanding of tumour evolution using next-generation-imaging sciences. We operate between the department of physics and the Cancer Research UK Cambridge Institute, and this gives us access to a community of optical physicists as well as cancer biologists, and a translational pipeline through to clinical application.

Build synergy

I'm a technology geek at heart. I love to dive into the underlying physics when I design optics and to explore the engineering of new devices; but, for me, the story is never complete without seeing these methods through to application. I feel that the research path I explore today is a synergy of all my experiences.

I am often asked what I learnt as I traversed this path from physics to medicine. The following are the key steps I had to take during my own personal development, which prepared me well to embrace my interdisciplinary career.

Learn the language. Just as with a foreign language, the fastest way to become fluent in the language of a new scientific discipline is to embed yourself inside it and practise regularly. For me, learning the language was the most important reason why I moved from physics to biochemistry for my first postdoc.

Get comfortable being uncomfortable. I went from being one of the most knowledgeable people in my research field at the end of my PhD to knowing less than most of the first-year PhD students in my new lab. For me, that meant I was doing something right, but you do have to be OK with taking that hit and spending time building a new skill set.

Don't forget that you bring unique skills. I had a lot to learn along my journey, but I brought useful skills at each stage. My quantitative mindset was valuable in designing imaging experiments, my engineering skills were useful for fixing broken equipment and my bottom-up physics approach to problems offered a fresh perspective.

Test biological-research hypotheses. Sometimes you'll bring a quantitative perspective that enables you to test a hypothesis or embark on a research topic that wasn't accessible before. Don't be afraid to ask the biological question yourself rather than waiting for your collaborators to provide an answer.



Physicist Sarah Bohndiek.

SARAH BOHNDIEK

> "A strong element of academic achievement is community recognition and networking."

Ask questions. I had to, and there are no stupid ones. I also attended introductory lectures for first-year undergraduates to fill in gaps in my knowledge. Going back to school and admitting what you do not know is crucial.

Embrace uncertainty. I continue to be amazed by the complexity of biology and hence the uncertainty in results. Unfortunately, it's likely that I'll never be as certain about anything in biology as I can be about new discoveries in physics.

Learn statistics. Having a quantitative background doesn't necessarily mean you really understand statistics. I had to dive back into statistics to design the best biological experiments.

Don't lose touch with your roots. Being able to continually innovate at the interface of physics and biology means staying grounded in the physical sciences, but this can be hard to do when you are in a biological department. To compete for faculty positions, you need to be able to teach your subject, so find ways to keep it fresh in your mind, as I did, for example, by teaching physics in undergraduate or graduate courses.

Do not blindly accept dogma. Challenging prevailing ideas in biology using your perspective can bring about revolutions. I greatly admire colleagues who have upturned decades of accepted dogma using quantitative methods that were not even considered by the biological community.

Perfect your pitch. Working across disciplines means you'll inevitably be talking at conferences and meetings to groups of scientific non-specialists; the best communicators perfect a balance of generalist and specialist material in presentations and make their language accessible to all.

Avoid equations during presentations. It's not just biologists who would prefer not to see them. Illustrate equations with graphs and visuals as much as possible. If you do have to include equations, write them on a board in real time so your audience can follow your guided explanation — don't just flash them up on the screen.

Find a good mentor. Finally, and most importantly, I would not be where I am today without my incredible mentors. When you move disciplines, you need people who can help you to integrate into a new community, show you how to navigate a new funding landscape and advise on expectations for junior faculty members, when the emphasis on publication and impact varies dramatically between disciplines.

The future

More unified training, encompassing physics and engineering applied to biology and medicine, will help to equip the next generation of researchers with the skills they need to operate across these disciplinary boundaries.

I still advocate for specialist training in one field as an undergraduate and then broadening research horizons gradually through advanced academic training.

A strong element of academic achievement is community recognition and networking; moving too rapidly between fields can make it challenging to build a reputation for excellence. Imaging is kind in this respect: regardless of how you form an image, there are many commonalities in the research that mean it is relatively easy to tell a story within your career and maintain connections along that journey.

Building that story and your profile will ultimately be crucial, no matter what path you choose to take.

Sarah Bohndiek leads an interdisciplinary team combining physicists, engineers, chemists and molecular biologists at the University of Cambridge, UK.

© 2020 Springer Nature Limited. All rights reserved.



Where I work
Martyn Poliakoff

Photographed for *Nature* by Leonora Saunders.

I usually apologize for the mess when people enter my office. There are dozens of periodic tables, hundreds of plastic bottles, pictures all over the walls, books going up to the ceiling and various old bits of apparatus that people wanted to throw out.

I'm a green chemist, interested in processes that generate as few hazardous by-products as possible. I'm part of a big project called Photo-Electro, funded by Britain's Engineering and Physical Sciences Research Council, that aims to minimize the amount of chemicals and solvents, and the number of processing steps, needed to construct complex molecules. For example, we study ways to use light and electricity to produce chemicals more efficiently.

But I'm perhaps better known as a popularizer of the periodic table. I have all sorts of objects inspired by it in my office: a large collection of periodic-table neckties, a now-empty box of periodic-table chocolates, and even a periodic table engraved on a strand of my own hair.

I also have a lot of toys that I play with: a dog toy that looks like carbon-60, various Rubik's cubes — I can't do Rubik's cubes, but they're fun to fiddle with — and silly things like a plastic frog that, when you squeeze it, produces fake spawn.

A big challenge for all scientists is to come up with new ideas. Workshops on creativity usually have toys lying around. Nobody is suggesting that a toy frog is necessarily going to inspire us to, say, use frog slime to produce a reaction. But I think that having lots of pictures, lots of toys hanging around is a good way of relaxing the mind and fostering innovation.

As for the plastic bottles, I had a long-running project trying to make the plastic in a greener way. We succeeded, but on a small scale, and the company decided it wasn't worth scaling up. Nonetheless, I learnt a lot, and the bottles are a good conversation starter. As it turns out, quite a lot of people enjoy being in my office.

Martyn Poliakoff is a green chemist at the University of Nottingham, UK. **Interview by Amber Dance.**

© 2020 Springer Nature Limited. All rights reserved.